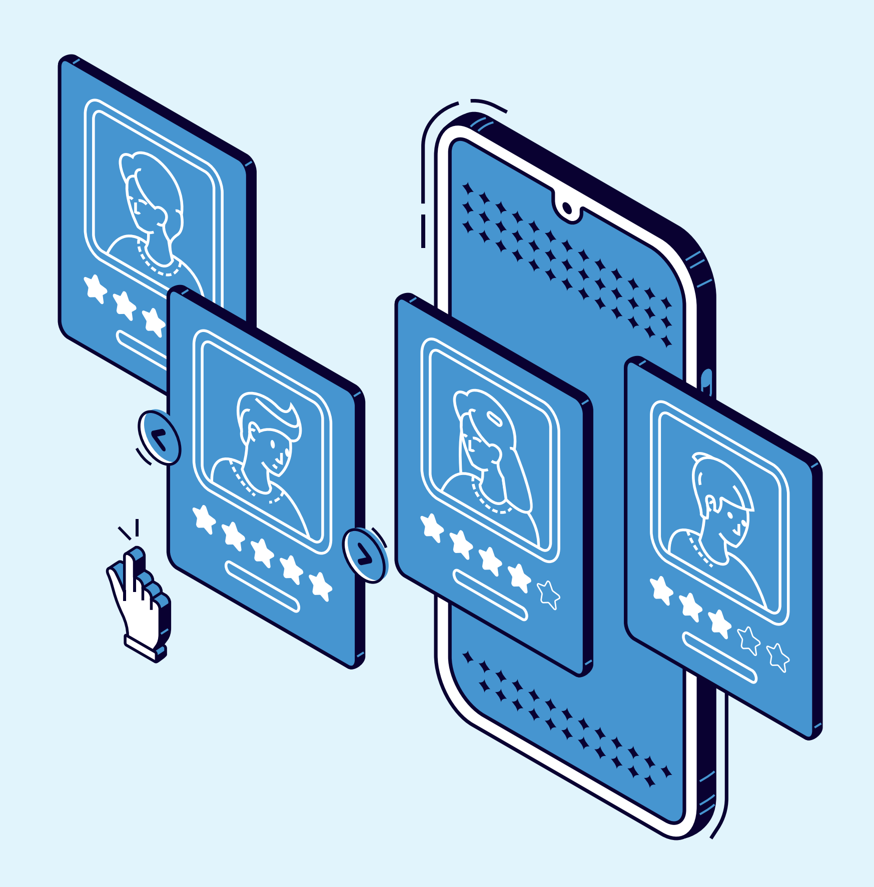
A Web-Based Approach to Measure Skill Mismatches and Skills Profiles for a Developing Country:

The Case of Colombia

Jeisson Cárdenas Rubio



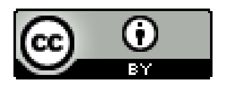




A web-based approach to measure skill mismatches and skills profiles for a developing country: the case of Colombia

Jeisson Cárdenas Rubio

CÁRDENAS RUBIO, J. *A web-based approach to measure skill mismatches and skills profiles for a developing country*: the case of Colombia [online]. Bogotá: Editorial Universidad del Rosario, 2020, 447 p. ISBN: 978-958-784-545-7. Available from. https://books.scielo.org/id/kt6p8. https://doi.org/10.12804/urosario9789587845457.




Universidad del
Rosario

A Web-Based Approach to Measure Skill Mismatches and Skills Profiles for a Developing Country





A Web-Based Approach to Measure Skill Mismatches and Skills Profiles for a Developing Country: The Case of Colombia

Abstract
Several interdisciplinary studies highlight imperfect information as a possible explanation of skill mismatches, which in turn has implications for unemployment and informality rates. Despite information failures and their consequences, countries like Colombia (where informality and unemployment rates are high) lack a proper labour market information system to identify skill mismatches and employer skill requirements. One reason for this absence is the cost of collecting labour market data.

Recently, the potential use of online job portals as a source of labour market information has gained the attention of researchers and policymakers, since these portals can provide quick and relatively low-cost data collection. As such, these portals could be of use for Colombia. However, debates continue about the efficacy of this use, particularly concerning the robustness of the collected data. This book implements a novel mixed-methods approach (such as web scraping, text mining, machine learning, etc.) to investigate to what extent a web-based model of skill mismatches can be developed for Colombia.

The main contribution of this book is demonstrating that, with the proper techniques, job portals can be a robust source of labour market information. In doing so, it also contributes to current knowledge by developing a conceptual and methodological approach to identify skills, occupations, and skill mismatches using online job advertisements, which would otherwise be too complex to be collected and analysed via other means. By applying this novel methodology, this study provides new empirical data on the extent and nature of skill mismatches in Colombia for a considerable set of non-agricultural occupations in the urban and formal economy. Moreover, this information can be used as a complement to household surveys to monitor potential skill shortages. Thus, the findings are useful for policymakers, statisticians, and education and training providers, among others.

Keywords: Job market (Colombia), tecnological innovations, hiring agents, data mining, web page development, data processing.

Medición de desajustes ocupacionales y perfiles de habilidades basados en datos de Internet: el caso de Colombia

Resumen
Varios estudios interdisciplinarios destacan la información imperfecta como una posible explicación del desajuste de habilidades, lo que a su vez tiene implicaciones para las tasas de desempleo e informalidad. A pesar de las fallas de información y sus consecuencias, países como Colombia (donde las tasas de informalidad y desempleo son altas) carecen de un sistema de información del mercado laboral adecuado para identificar los desajustes de habilidades y los requisitos de habilidades de los empleadores. Una de las razones de esta ausencia es el costo de recopilar datos sobre el mercado laboral.

Recientemente, el uso potencial de portales de empleo en línea como fuente de información sobre el mercado laboral ha atraído la atención de investigadores y legisladores, ya que estos portales pueden proporcionar una recopilación de datos rápida y de costo relativamente bajo. Como tal, estos portales podrían ser útiles para Colombia. Sin embargo, continúan los debates sobre la eficacia de este uso, particularmente en lo que respecta a la solidez de los datos recopilados. Este libro implementa un enfoque novedoso de métodos mixtos (como *web scraping*, minería de texto, aprendizaje automático, etc.) para investigar hasta qué punto se puede desarrollar un modelo basado en la web de desajustes de habilidades para Colombia.

La principal contribución de este libro es demostrar que, con las técnicas adecuadas, los portales de empleo pueden ser una fuente sólida de información sobre el mercado laboral. Al hacerlo, también contribuye al conocimiento actual al desarrollar un enfoque conceptual y metodológico para identificar habilidades, ocupaciones y desajustes de habilidades utilizando anuncios de empleo en línea, que de otra manera serían demasiado complejos para ser recopilados y analizados por otros medios. Al aplicar esta metodología novedosa, este estudio proporciona nuevos datos empíricos sobre el alcance y la naturaleza de los desajustes de habilidades en Colombia para un conjunto considerable de ocupaciones no agrícolas en la economía urbana y formal. Además, esta información se puede utilizar como complemento de las encuestas de hogares para monitorear la posible escasez de habilidades. Por lo tanto, los hallazgos son útiles para los encargados de formular políticas, los estadísticos y los proveedores de educación y capacitación, entre otros

Palabras clave: Mercado laboral (Colombia), innovaciones tecnológicas, agentes de empleo, minería de datos, desarrollo de páginas web, procesamiento de la información.

Suggested citation / Citación sugerida
Cárdenas Rubio, Jeisson. 2020. *A Web-Based Approach to Measure Skill Mismatches and Skills Profiles for a Developing Country: The Case of Colombia*. Bogotá, D. C.: Editorial Universidad del Rosario.
https://doi.org/10.12804/urosario9789587845457

A Web-Based Approach to Measure Skill Mismatches and Skills Profiles for a Developing Country:
The Case of Colombia

Jeisson Cárdenas Rubio

Cárdenas Rubio, Jeisson
A web-based approach to measure skill mismatches and skills profiles for a developing country: the case of Colombia / Jeisson Cárdenas Rubio. – Bogotá: Editorial Universidad del Rosario, 2020.

XXII, 447 páginas
Incluye referencias bibliográficas.

1. Mercado laboral – Innovaciones tecnológicas – Colombia. 2. Agentes de empleo. 3. Web – Minería de datos. 4. Desarrollo de páginas web. 5. Procesamiento de la información I. Cárdenas Rubio, Jeisson. II. Universidad del Rosario. III. Colombia Científica. Conocimiento global para el desarrollo. Alianza Efi. Economía formal e inclusiva. IV. Título.

331.1201132 SCDD 20

Catalogación en la fuente – Universidad del Rosario. CRAI

DJGR Octubre 16 de 2020

Hecho el depósito legal que marca el Decreto 460 de 1995

Editorial Universidad del Rosario
Carrera 7 No. 12B-41, of. 501
Tel: 297 0200 Ext. 3112
https://editorial.urosario.edu.co/

Primera edición: Bogotá, D. C., 2020

ISBN: 978-958-784-544-0 (impreso)
ISBN: 978-958-784-545-7 (ePub)
ISBN: 978-958-784-546-4 (pdf)
https://doi.org/10.12804/urosario9789587845457

Coordinación editorial: Editorial Universidad del Rosario
Corrección de estilo: Erika Tanacs
Diseño de cubierta: Juan Ramírez
Diagramación: William Yesid Naizaque Ospina
Impresión: Xpress. Estudio Gráfico y Digital SAS

Impreso y hecho en Colombia
Printed and made in Colombia

Contents

Appendix

List of Figures

List of Tables

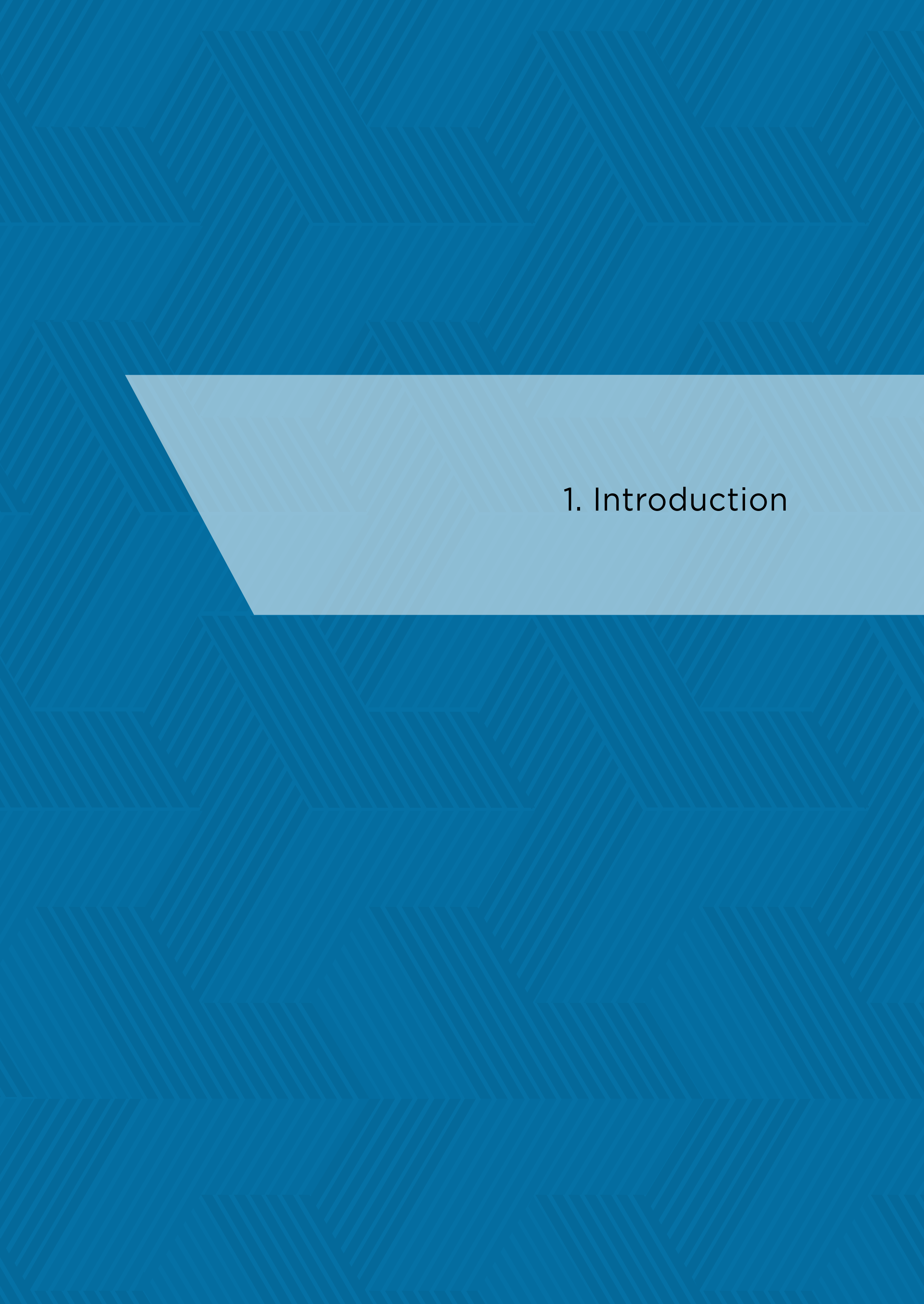
Acronyms and Abbreviations

AM	Metropolitan Areas (for its acronym in Spanish)
API	Application Program Interface
APL	A Programming Language
ASP	Active Server Pages
BBVA	Banco Bilbao Vizcaya Argentaria
BPM	Business Process Management
CASCOT	Computer Assisted Structured Coding Tool
CE	Cambridge Econometrics
CEDEFOP	European Centre for the Development of Vocational Training (for its acronym in Spanish)
CEPAL	Comisión Económica para América Latina y el Caribe
CERES	Regional Centres of Higher Education (for its acronym in Spanish)
CNC	Computer Numerical Control
CONPES	Consejo Nacional de Política Económica y Social
CSS	Cascading Style Sheet
CVTS	Continuing Vocational Training Survey
DANE	Departamento Administrativo Nacional de Estadística
DEEWR	Australian Department of Education, Employment and Workplace Relations
DfE	Department for Education
DG	Directorate-General
EAP	Economically Active Population

EB	Exabyte
ECLAC	Economic Commission for Latin America and the Caribbean
EEA	European Economic Area
EFCH	Encuesta de productividad y formación de capital humano
ESCO	European Skills, Competences, Qualifications and Occupations
ESS	Employer Skills Survey
EU	European Union
FILCO	Fuente de Información Laboral de Colombia
GDP	Gross Domestic Product
GEIH	Gran Encuesta Integrada de Hogares
HSEQ	Health, Safety, Environment & Quality
HTML	Hypertext Markup Language
IALS	International Adult Literacy Survey
ICT	Information and Communications Technology
IDB	Interamerican Bank of Development
IER	Warwick Institute for Employment Research
ILO	International Labour Organization
IP	Internet Protocol
ISCO	International Standard Classification of Occupations
ISIC	International Standard Industrial Classification of All Economic Activities
ISO	International Organization for Standardization
IT	Information Technology
LASSO	Least Absolute Shrinkage and Selection Operator
LEFM	Local Economy Forecasting Model
LFS	Labour Force Survey
LTDA	Limitada
MAC	Migration Advisory Committee
MEN	Ministerio de Educación Nacional de Colombia
N&E	New and Emerging (Occupations)
NIF	Normas de Información Financiera
NIIF	Normas Internacionales de Información Financiera
NOS	National Occupational Standards
NQF	National Qualifications Framework
OECD	Organisation for Economic Co-operation and Development
OEI	Organización de Estados Iberoamericanos

OLS	Ordinary Least Squares
O*NET	Occupational Information Network
ONS	Office for National Statistics
OSP	Occupational Skills Profiles
OVATE	Skills Online Vacancy Analysis Tool
PHP	Hypertext Preprocessor
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
RSPO	Roundtable on Sustainable Palm Oil
RUES	Registro único empresarial
SENA	Servicio Nacional de Aprendizaje
SEO	Search Engine Optimization
SIC	Standard Industrial Classification
SMEs	Small and Medium-Sized Enterprises
SMMLV	Salario mínimo mensual legal vigente
SNIES	Sistema Nacional de Información de Educación Superior
SNPP	Sub-National Population Projections
SOC	Standard Occupational Classification
SQA	Software Quality Assurance/Advisor
SQL	Structured Query Language
SST	System Support Team
SSTA	Gestión en seguridad, salud en el trabajo y ambiente
STEP	Skills Measurement Program
SVM	Support Vector Machine
TAT	Store-to-store (for its acronym in Spanish)
TVET	Technical and Vocational Education and Training
UAESPE	Unidad Administrativa Especial del Servicio Público de Empleo
UK	United Kingdom
UKCES UK	Commission for Employment and Skills
US	United States
VET	Vocational Education and Training
XML	Extensible Markup Language





1. Introduction

This book studies how, and to what extent, a web-based system to monitor skills and skill mismatches could be developed for Colombia based on information from job portals. More specifically, this document seeks to answer the following questions: 1) How can information from job portals be used to inform policy recommendations? And, in order to address two of the major labour market problems in Colombia, which are high unemployment and informality rates, 2) to what extent can information from job portals (unsatisfied demand) and national household surveys (labour supply) be used together to provide insights about skill mismatch issues in a developing economy?

Consequently, this book investigates the challenges, advantages, and limitations of collecting information from job portals and proposes a framework to test this information's validity for economic analysis. It conducts an innovative labour market analysis and develops indicators based on updated and robust labour demand (job portal) and labour supply (household survey) information to tackle skill mismatches, extending thus the use of novel sources of information to yet unexplored areas in the existing labour economics literature.

By doing so, this study makes conceptual, methodological, and empirical contributions to the ongoing debate in economics about the use of information from job portals for labour demand analysis. The main conceptual contribution consists of demonstrating that the concept and sources of Big Data (in this case, job portal sources) can provide consistent results to orient public policies (see Chapters 7 to 9). This document also demonstrates that, with the proper techniques, information from job portals can fulfil conceptual requirements to be considered as high-quality data for labour market analysis (see Chapters 4 and 10).

The main methodological contribution is the development of a detailed framework and methods to collect, clean, and organise (i.e. web scraping, occupation and skill identification, etc.) vacancy data, which allows testing and analysing this source of information for consistent labour market insights.

Specifically, this book contributes to the methodology of processing information from job portals for public policy advise by: 1) discussing different criteria (volume, website quality, and traffic ranking) to select the most relevant and trustworthy job portals in order to collect vacancy information (Chapter 5); 2) providing a detailed explanation about Big Data techniques (web scraping) and the challenges they pose for automatically collecting job advertisements from job portals (Chapter 5); 3) applying mixed-methods approaches (text mining, word-based matching methods, etc.) to standardise information collected from different job portals into a single database for statistical analysis (Chapter 6); 4) implementing and extending a mixed-methods approach (stop words, stemming, extensions of a machine learning algorithm, etc.) in order to identify skills and occupations in online job announcements (Chapter 6); 5) and, importantly, using this extended mixed-methods approach (e.g. a skills dictionary to identify skill patterns) to find new or specific skills and occupations in the Colombian labour market, which would otherwise be complex to identify via other means (e.g. household surveys) (Chapter 6).

Moreover, the book proposes a (n-gram-based) method to reduce duplication issues (as information is collected from different job portals, some job advertisements can be repeated) and a (Lasso) method to impute missing values, such as education and wages (Chapter 6). Consequently, by implementing and extending novel mixed methods, 6) this document improves data collection and helps to understand methodological changes to collect and organise information from job portals.

As a product of the above methods, a vacancy database was consolidated for the period between January 1, 2016 and December 31, 2018 (Chapter 7). In addition, this document makes further methodological contributions by 7) proposing a framework to evaluate the internal (consistency) and external (representativeness) validity of this vacancy database. To test internal validity, a statistical comparison was conducted between variables, such as wages, occupations, education, etc., to understand biases, errors, and inconsistencies within the database. The evaluation of external validity was particularly challenging because countries like Colombia do not have vacancy censuses (or anything similar) to compare information collected from job portals. Despite several obstacles, this book provides and applies a methodology framework to evaluate the vacancy database. It implements a detailed comparison between official

information available in the country (i.e. household surveys) and vacancy data results, such as vacancy, employment, new hires, unemployment, occupational structures and their dynamics over the study period. This comparison enables the understanding of possible biases (e.g. over/underrepresentation of certain occupational groups) in the vacancy database (Chapter 8).

Based on the validation results, another methodological contribution of this document is 8) proposing and estimating skill mismatch measures that consider the advantages and limitations of job portals and household surveys. Specifically, the study demonstrates how household surveys can be combined with vacancy data to produce relevant (volume- and price-based) skill shortage indicators, such as percentage change in unemployment by sought occupation, percentage change in median real hourly wage, among others. Importantly, 9) this book makes an important contribution to the discussion about skill mismatch measures by considering informality. As will be discussed in Chapter 9, informality is a signal of labour market imbalance. A considerable portion of employment growth might be explained because people cannot find a formal job and have to choose informal jobs. Thus, skill shortage indicators need to control for informality to avoid misleading results.

Based on the above methodology, this book also makes relevant empirical contributions by providing a detailed labour market analysis that reveals important characteristics of the Colombian labour demand (e.g. demanded skills and occupational trends). Importantly, it determines skill mismatches (i.e. skill shortages) in Colombia based on information from job portals and household surveys. Specifically, the analysis of the vacancy database evidences that 1) data collected from job portals are representative of a considerable set of non-agricultural, non-governmental, non-military, and non-self-employed ("business owners") occupations; 2) most of the vacancies in Colombia correspond to middle- and low-skilled occupations (such as "Sales demonstrators"); 3) in alignment with the most demanded occupations, the most demanded skills are "Customer service," "Work in teams," etc.; and, most importantly, 4) information from job portals can be used to identify new or specific job titles (e.g. "TAT vendors," "Picking and packing assistants," etc.) and skills (e.g. "Siigo," "*Perifoneos*," etc.) for the Colombian context.

Based on the advances made towards homologating vacancy and household survey information (e.g. coding both databases according to ISCO-08),

a comprehensive analysis of labour demand and supply information is conducted at the occupational level (Chapter 9), for the first time in Colombia. Another important contribution of this analysis consists of 5) showing in detail population groups with higher (lower) informality and unemployment rates. For instance, domestic cleaners and helpers and motorcycle drivers face the highest informality, while environmental engineers and geologists and geophysicists face the highest unemployment rate in the country. In addition, 6) it also estimates skill shortages using job portals and vacancy information. For instance, it evidences that 30 occupations show signals of skill mismatches, while indicating that Structured Query Language (SQL), database management, and JavaScript are the most demanded skills for one of those occupation groups ("Web and multimedia developers").

Briefly, skill mismatches arise when there is a misalignment between the demand and supply of skills in the labour market (UKCES 2014). As will be discussed in Chapters 2 and 3, numerous multidisciplinary studies have pointed out the importance of these phenomena in labour market outcomes, such as unemployment and informality, among others. Skill mismatches can occur in the job search process (e.g. skill shortages) or in the workplace (e.g. skill gaps). Given that the term "skill mismatches" encompasses different dimensions and considering available data to analyse an economy such as Colombia (i.e. job portals and household surveys), this book focuses on studying skill shortages. This concept refers to issues that arise in the job searching process when jobseekers do not have the proper skills required in vacancies posted by employers (Green, Machin, and Wilkinson 1998).

A proper labour market analysis system to identify possible skill shortages and current employer skill requirements is paramount for a country such as Colombia with high and persistent unemployment and informality rates (DANE 2017a). According to the Colombian statistics office (National Administrative Department of Statistics; DANE for its acronym in Spanish), in the last two decades unemployment and informality rates were around 12.5% and 49.4%, respectively. A vast number of factors, such as rigid wages, comparatively high non-wage costs, etc., could explain these labour market outcomes. However, as will be discussed in Chapters 2 and 3, theoretical and empirical evidence shows that mismatches between demanded skills and those offered is a main cause of unemployment and increased informality rates in Colombia (Álvarez and

Hofstetter 2014; ManpowerGroup, n.d.; Arango and Hamann 2013). Workers, the government, as well as education and training providers are not properly anticipating employer requirements. Consequently, the labour supply lacks skills in relation to what employers are demanding in order to fill their vacancies.

Despite evidence that suggests that there is a high incidence of skill shortages in the Colombian labour market, education and training providers, workers, and the government can do little to reduce imperfect information regarding human capital requirements due to a lack of proper information to develop well-orientated decisions and public policies (González-Velosa and Rosas-Shady 2016). On the one hand, the cost of conducting household or sectoral surveys (traditional sources of information) is relatively high in terms of resources and time. On the other hand, these data sources usually fail to provide detailed and updated information about skills and occupational requirements. These issues have discouraged countries (especially those with low budgets) from collecting information on and analysing human capital needs.

For instance, the Colombian office for national statistics (DANE) periodically conducts household and sectoral surveys that provide valuable insights about the characteristics of the Colombian workforce, job training, selection and hiring practices, productivity, etc. However, due to sample constraints and the relatively high operational cost of conducting these surveys (e.g. the job of interviewers and statisticians, etc.), the data collected do not convey detailed information about employer requirements—the occupational structure demanded—nor about the skills required for each position. Thus, the characteristics and dynamics of labour demand remain relatively unknown.

Consequently, to fill these critical information gaps, it is vital to seek new ways of analysing labour demand that can consistently complement existing surveys (e.g. household surveys). Big Data have become a trendy field because it deals with the analysis of large data sets, in real time, from different sources of information (Edelman 2012; Reimsbach-Kounatze 2015). Using job portals and Big Data techniques to analyse employer requirements constitutes an alternative that has attracted the attention of researchers and policymakers. Employers post a considerable number of vacancies on online job portals along with detailed candidate requirements (job title, wages, skills, education, experience, etc.), which provides quick access to a large amount of relevant information for the analysis of labour demand. This online data can provide key

insights about labour demand that previously were not accessible for proper analysis (Kureková, Beblavy, and Thum 2014).

Collecting, processing, and analysing information from job portals through reliable and consistent statistical processes is challenging because data are dispersed across different websites and the information is not categorised or standardised for economic analysis. Additionally, the discussion regarding the use of Big Data sources, such as job portals for labour market analysis, is flawed (Kureková, Beblavy, and Thum 2014). Different authors have used and derived conclusions from job portal data without considering in detail the possible biases and limitations of this information (e.g. Backhaus 2004; Kureková, Beblavy, and Thum 2016; Kennan et al. 2008). Like any other source of data, information from job portals has biases and limitations. For instance, given the type of internet users, among other data quality issues, job portals are unlikely to be representative of the whole economy or a specific sector, or they might not reflect real trends in labour demand. The lack of debate concerning data validity has affected the credibility of job portals as a consistent and useful resource for labour market analysis.

A conceptual and methodological framework is required in order to use vacancy data and to properly address issues such as skill mismatches. Therefore, this book seeks a better understanding about the use of new sources such as job portals to analyse the labour market (skill mismatches) in a developing country such as Colombia. This study responds to the need to develop a more efficient way to collect and analyse information about labour demand and skills in order to identify potential skill shortages. This kind of work supports the design of national skills strategies, while enhancing the capacity of governments to develop public policies to tackle current skill mismatches (Cedefop 2012a).

To this end, this book is structured as follows: Chapter 2 discusses the concepts and theoretical framework used in this document to analyse labour market based on the information found on online job portals. First, this chapter introduces basic conceptual and statistical definitions for labour demand (e.g. job vacancies) and labour supply (e.g. unemployed and employed workers). Second, given that a considerable share of the population in Colombia works in irregular market conditions, this chapter discusses what is understood in the academic literature by informality. Furthermore, the concept of skills and different ways to measure them for economic analysis are examined. Subsequently,

the previously mentioned definitions are used to describe the dynamics of the labour market and its main outcomes, such as unemployment, wages, etc., under the assumption of perfect competition (e.g. assuming that companies and workers are perfectly informed about the quality and the price of "labour"). Nevertheless, the assumptions of perfect competition are unrealistic given that workers are usually not perfectly aware of employer skill requirements; similarly, this model is not an appropriate theoretical framework for economies such as Colombia (Garibaldi 2006). Based on a model with imperfect information (which seems more appropriate to describe Colombian labour market outcomes), Chapter 2 explains how skill mismatches can arise, as well as their consequences for informality and unemployment rates (Bosworth, Dawkins, and Stromback 1996; Reich, Gordon, and Edwards 1973; Stiglitz et al. 2013). This framework highlights that information failures might be one of the leading causes of high unemployment and informality rates. Thus, actions to decrease these information failures (such as the use of job portals) will considerably improve people's employability.

Chapter 3 presents evidence that skill shortages, unemployment, and informality are high-frequency phenomena in Colombia (DANE 2017a; ManpowerGroup, n.d.; Arango and Hamann 2013). Moreover, it outlines how the government, as well as education and training providers, etc., face severe difficulties to tackle these issues due to the lack of a proper system to identify skills in demand and possible skill shortages (González-Velosa and Rosas-Shady 2016). First, the chapter describes the main characteristics of the Colombian labour market, such as unemployment, informality, etc., and their evolution during the last two decades. In addition, it provides a general description of the socio-economic characteristics of the labour force and—based on the little information available—the labour demand. Second, it evidences a high incidence of skill shortages in Colombia and their possible implications for labour market outcomes. It is argued that workers, education and training providers, as well as the government can do little to address these issues given the lack of proper information to monitor and identify employer requirements and possible skill shortages at the occupational level. Subsequently, the chapter presents an overview of the Colombian labour market focused on unemployment, informality, and skill shortages, and highlights the need for detailed information to adequately address these issues.

In Chapter 4, the concept of Big Data is introduced, with its advantages and limitations outlined for a labour market analysis. Moreover, this chapter explains why traditional statistical methods, such as household or sectoral surveys, encounter difficulties in providing detailed information about the labour market. First, it defines Big Data according to three properties: volume, variety, and velocity (Laney 2001). Then, it discusses the problems of traditional statistical methods, such as sample or survey design, that constrain labour market analysis in terms of occupations and skills (Kureková, Beblavy, and Thum 2014; Reimsbach-Kounatze 2015). Given these information gaps, the potential use of Big Data sources to complement labour market analysis is discussed, with a special focus on job portals and their possible application to tackle skill shortages. Subsequently, this chapter explains the limitations and caveats to be considered when online vacancy data are used for economic analysis. Furthermore, it emphasises the differentiating features of this book, compared to other ongoing studies.

Once the conceptual framework and the need for information and analysis to address skill shortages are established, Chapters 5 and 6 present a comprehensive methodology to systematically collect and standardise vacancy information from job portals. Chapter 5 describes available information that can be collected from Colombian job portals. Then, it proposes criteria to consider the volume of information on each job portal, as well as each website's quality and traffic ranking to select the most important and reliable job portals for an analysis of the labour demand in Colombia. Subsequently, Chapter 5 describes the methodology (web scraping) and different challenges to automatically and rapidly collect a massive number of online job vacancies. The chapter also explains the methods that can be used to homogenise variables such as education and experience and to consolidate information from job portals into a single database.

Next, Chapter 6 illustrates the methods and challenges involved in standardising two of the most relevant variables for the economic analysis of the labour market: skills and occupations. Furthermore, this chapter examines the issues of duplication and missing value, which are some of the main concerns when analysing information from job portals. First, the chapter develops a method to automatically identify skill patterns in job vacancy descriptions based on international skill descriptors and text mining. Then, it proposes

and applies a novel mixed-method approach (software classifiers and machine learning algorithms) to properly classify job titles into occupations. Third, as an employer might advertise the same job many times on the same job portal or on different job portals, the chapter identifies and minimises the issue of duplication. It also explains how missing values were imputed for the "educational requirement" and "wage offered" variables (which are relevant to test the validity of the vacancy database and to analyse labour demand) by using predictors such as occupation, city, and experience requirements. As a result of the above methods, a Colombian vacancy database is generated in Chapter 6 to be tested and analysed to address skill shortage issues.

Subsequently, a comprehensive descriptive analysis of the Colombian labour demand is conducted in Chapter 7. First, the analysis describes the selected job portals, as well as their geographic distribution, in order to build the mentioned vacancy database. Second, it provides a detailed descriptive analysis of the labour demand for skills in Colombia, such as education, occupational structure, potential new occupations, and skills and experience requirements. This description reveals characteristics of the labour demand that were unknown prior to this study. Third, this chapter examines the most notable labour demand trends by occupation: those with higher demand, those with a higher increase, and occupations for which demand has decreased over time. Finally, it describes the distribution of wages offered by employers and other secondary characteristics of the vacancy database, such as contract types and the duration of vacancies.

Although this descriptive analysis might have considerable implications for policymakers and researchers, these results do not provide enough evidence about the validity or reliability of vacancy data to address skill shortages and their consequences. As is the case with data collected by other methods (e.g. surveys), information collected from job portals have limitations that affect interpretation (Chapter 4). Consequently, there is a critical need to assess the validity of the vacancy database to be sure of what it can tell us about labour demand. Thus, Chapter 8 performs extensive internal and external validity tests on the vacancy database (Henson 2001; Rasmussen 2008). First, it evaluates internal validity (consistency of the variables within the same database) via cross-tabulations and wage distribution analysis. Second, it tests external validity (representativeness) of the online vacancy information. This examination

requires a comparison of the vacancy database results against other sources of information (e.g. household surveys). To do so, this book re-categorises occupations from Colombian household surveys to create updated occupational classifications that are compatible with occupational categories in the vacancy database.

After completing the described homologation, a "traditional" test is conducted by comparing the occupational structures of supply and demand. However, given the limitation of the "traditional" test, further tests are carried out to investigate the external validity of the database. Specifically, the wage distribution of labour demand and information on supply are examined to perform a relevant comparison of time series between jobs in demand, employed and unemployed individuals in the total workforce, as well as the extent of new hires (replacement demand and employment growth) by major occupational groups. These detailed tests provide information about the advantages and limitations of the vacancy database for a labour demand and skill mismatch analysis.

Once the advantages and limitations of the data are established, Chapter 9 proceeds to develop a system to identify possible skill shortages and address labour supply according to employer requirements in Colombia. First, the chapter provides a detailed description of the Colombian labour market panorama (formally or informally employed, as well as unemployed) at the occupational level. Second, it combines Colombian household survey information and the vacancy database to estimate a Beveridge curve and a set of eight (volume- and price-based) macro-indicators to identify possible skill shortages. This chapter also highlights the importance of controlling for informality when building skill mismatch indicators in a context such as Colombia. Occupations might exist with relatively low unemployment rates, but also with a relatively high informality rate, or vice versa. Accordingly, increases in the number of workers in certain occupations—for instance, those characterised by relatively low unemployment rates—might increase informality rates. Therefore, this document advises policymakers and training providers to be aware of this relevant labour market duality when providing and promoting skills. Furthermore, this chapter shows how detailed information from vacancies (job descriptions) can be used to monitor labour demand trends for skills, as well as to update occupational classifications according to current employer requirements.

Finally, Chapter 10 summarises the relevant conceptual, methodological, and empirical contributions of the book, while opening a debate on the use of novel sources of information (job portals) to fill information and analysis gaps regarding the labour market. Thus, this chapter highlights the implications of the findings for national statistics offices, policymakers, education and training providers, and career advisers. Additionally, it points out the limitations of the study and illustrates new avenues of enquiry for future research.

This comprehensive and detailed methodological and conceptual framework alongside empirical findings presents important evidence about the advantages and limitations of job portals for their use in economic analysis. It provides a basis to develop a consistent skill shortage monitoring system that can be beneficial for different countries when adopted.

2. The Labour Market and Skill Mismatches

2.1. Introduction

The labour market can be defined as a "place" (not necessarily a physical place) where employers ("demand") and workers ("supply") interact with each other. The dynamics of this labour market are relevant for an economy as they determine different socio-economic outputs, such as productivity, unemployment, wages, and poverty, among others. Provided that the labour market influences various outcomes and different disciplines address these issues (e.g. sociology, economy, etc.), this chapter discusses labour market definitions and explains the theoretical framework adopted throughout this book to analyse labour demand based on the information found on online job portals.

The second section of this chapter explains what is understood by labour demand and labour supply in the academic literature on economics, and possible ways to statistically measure these concepts. Moreover, it defines and highlights informal economy as a key issue, especially in Latin American countries like Colombia. Subsequently, the concept of skills is introduced, and its possible implications for unemployment and informal economy are explained. With these basic definitions outlined, the third section of the chapter describes the Colombian labour market and its main outcomes, such as unemployment, wages, etc., under the assumption of perfect competition.

However, the assumptions of perfect competition are substantial and might not be appropriate for different economies such as the Colombian economy. Consequently, it is necessary to consider labour market failures—for example, imperfect information—that might appropriately explain the comparatively high rates of informal economy and unemployment levels in Colombia. Thus, the fourth section of this chapter focuses on explaining how imperfect information might increase skill mismatches and, consequently, create labour market segmentation between formal and informal workers along with a comparatively high unemployment rate, proposing thus that information failures might be one

of the leading causes of high unemployment and informality rates, especially in developing countries like Colombia.

2.2. Basic definitions

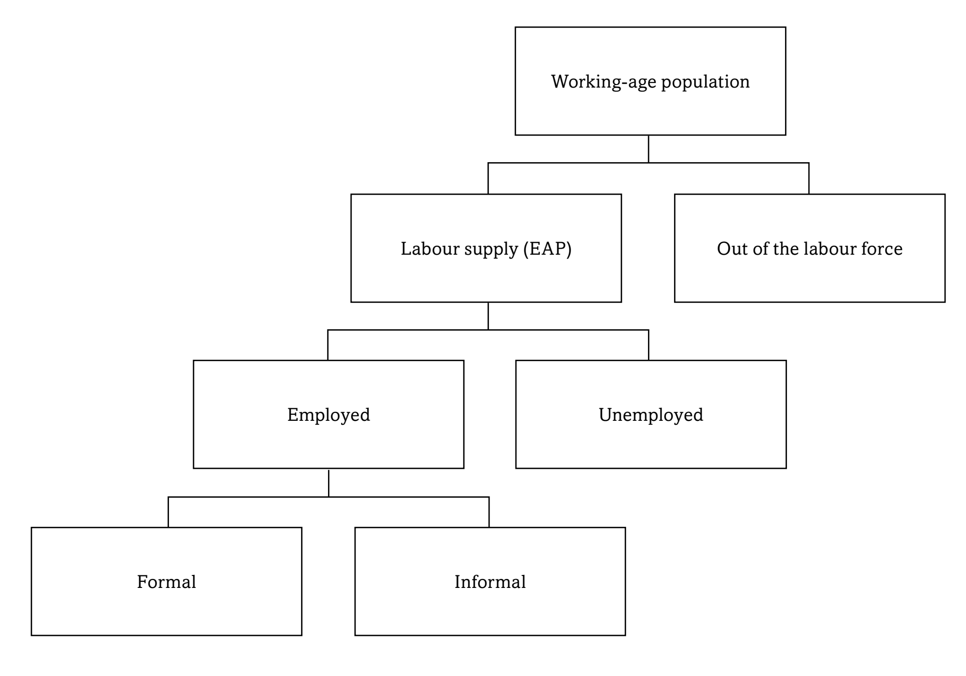
Comparable to other markets (e.g. financial markets, physical consumer markets, etc.), the labour market is composed of demand and supply (Cahuc, Carcillo, and Zylberberg 2014). The merchandise to be exchanged consists of "labour services" that represent human activities (distinguished by numbers of workers or hours of work); these human activities are one of the inputs in the production of goods and services (ILO 2018). Consequently, the dynamics between demand and supply have various implications for a range of individuals, for instance, for people with different characteristics (i.e. skills), employers who create job offers with certain requirements, and government institutions, among others. Thus, this section explains who form the labour demand and labour supply (e.g. unemployed, formal, and informal workers), as well as the relevance of skills in labour market outcomes.

2.2.1. Labour supply

In a basic economic model, people or households possess a limited quantity of "labour" that they can offer in the labour market in order to have an income to acquire goods and services (Cahuc, Carcillo, and Zylberberg 2014). Therefore, the labour supply or labour force is composed of people who offer their "labour." As shown in Figure 2.1, the labour supply or economically active population (EAP) is composed of 1) people who do not have a job but are looking for one (unemployed) and 2) people who are part of the working-age population hired by employers (employed) or are self-employed (ILO 2010). For statistical purposes, according to the International Labour Organization (ILO), all working-age people who did not participate in the production of goods and services for at least one hour in the reference week (one week before the survey is conducted) because they did not need to, could not or were not interested in earning a labour income, are considered out of the labour force (or inactive) (ILO 2010). An unemployed individual is a person without work

who has sought a job during the last four weeks and is available for work within the next fortnight; or who is currently without a job but has accepted a job to start in the next fortnight. An employed individual is employed when he/she worked for at least one paid or unpaid hour in the reference week. These employed and unemployed individuals are considered as labour force (EAP).

Figure 2.1. **Labour market structure**



Source: Author's elaboration.

2.2.2. Labour demand

In contrast, companies or establishments require "labour services" as an input to produce goods and services in the private and public sectors. Consequently, labour demand refers to the demand for workers (or hours of work) in an economy. This demand consists of the level of employment (satisfied labour demand) plus the number of available job vacancies, which equates to the labour required but not filled by an employee over a certain period of time (unsatisfied labour demand or unmet demand) (Farm 2003; Williams 2004).

In this sense, a job vacancy is defined as a "paid post that is newly created, unoccupied, or about to become vacant:

a) for which the employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned; and
b) which the employer intends to fill either immediately or within a specific period" (Eurostat 2017).

Therefore, the total number of vacancies in an economy is determined by the number of unfilled job openings and, additionally, the number of jobs that are temporarily filled by internal substitutes (Farm 2003).

To summerize this subsection, the classic economic model (Cahuc, Carcillo, and Zylberberg 2014) describes the labour market in the following way: people (or households) offer a certain quantity of their "labour" at a certain labour price level (wages) in order to generate income and acquire different goods and services available in other markets. At the same time, establishments in this model require a certain quantity of "labour" at a certain labour price level (wages) to produce goods and services, and while some workers have a job and are employed, others are looking for one and are unemployed. Nevertheless, as shown in Figure 2.1, the fact that people are employed does not imply that they are working in regulated and good working conditions (e.g. informal economy).

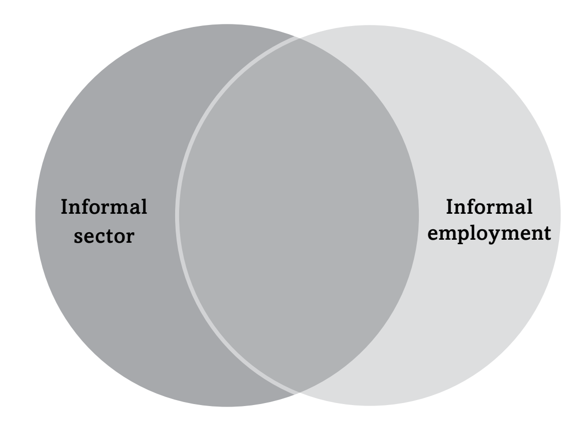
2.2.3. Informal economy

To measure informal economy, the ILO (2003) recommends making a distinction between informal sector and informal employment. On the one hand, the informal sector is an enterprise-based definition, which considers people working in units that have "informal" characteristics regarding their unregistered and/or unincorporated legal status, small size, non-registration of their employees, lack of formal labour relations and bookkeeping practices, as well as under-payment/non-payment of taxes, among others. On the other hand, informal employment is a job-based definition and covers individuals whose main job lacks basic legal and social protections (or employment benefits); for

example, lack of social protection, no income taxation, and so forth. It is necessary to clarify that none of these informal economy concepts refer directly to underground, illegal, and non-market production. These kinds of activities belong to illegal economy and (usually) are difficult to measure with standard labour market surveys such as household or sectoral surveys (Perry et al. 2007).

The above definitions of informal sector and informal employment highlight different aspects of an informal economy and can be used for various public policy objectives, such as payroll taxes, social protection, among others (ILO 2003). Consequently, it is possible that people work informally for enterprises that operate in the formal economy or workers might have formal jobs (e.g. with social security) in enterprises in the informal sector (see Figure 2.2).

Figure 2.2. **Composition of informal economy**



Source: Author's elaboration.

Based on ILO recommendations, in national household surveys, countries like Colombia consider the following individuals to be informal workers: private employees and workers in establishments, businesses or companies that occupy up to five people (in all their agencies and branches), including the employer and/or partner; unpaid family workers; domestic employees; self-employed workers, except independent professionals, while government employees are excluded from this definition (Hussmanns 2004). Evidently, Colombia's household surveys classify informal workers according to the concept of informal

sector.[1] As mentioned by Freije (2002), despite there is no consensus on how to measure informality, most of the researchers in Latin America rely on the firm size approach to measure this phenomenon. Indeed, the ILO (2019) reported that 13 out of 18 countries surveyed in Latin America[2] include firm size as a criterion when defining the informal sector.

However, this way of measuring informality has some limitations. As mentioned above, informal employees may be formally working in large factories and, in consequence, the way in which Colombia measures informality might underestimate the phenomenon (ILO 2012). However, using a measure employed by the DANE to calculate informality via social security contributions (pensions) and firm size, Bernal (2009) found that—at least for the Colombian case—the size of the informal sector is remarkably similar between social security and firm size informality measurements. The same author found that workers who pay social security contributions (pension and/or health) are less likely to belong to small firms. In addition, the ILO (2011) studied 47 medium- and low-income countries and concluded that almost all workers employed by the informal sector are also in informal employment.

Another concern with the firm size criterion is that all self-employed workers might be considered as informal workers. According to monthly labour market figures released by the DANE,[3] around 80% of self-employed workers and around 20% of salaried workers are informal. Consequently, the firm size informality definition tends to be correlated with self-employment, but the relation is not one to one. Additionally, for 14 Latin American countries,[4] Perry et al. (2007) demonstrated that firm size (among other variables, like

---

1 Even though official informality statistics are based on the concept of the informal sector, it is possible to calculate informality using an informal employment approach (e.g. pension and/or health contributions, among other benefits). Moreover, Colombia excludes agricultural activities from its official informal sector statistics, since including such activities requires developing a more robust definition, especially regarding jobs held by own-account workers and members of producer cooperatives in the agricultural industry (ILO 2003).

2 Argentina, the Plurinational State of Bolivia, Brazil, Colombia, Costa Rica, Dominican Republic, Guatemala, Guyana, Honduras, Jamaica, Mexico, Panama, Paraguay, Peru, El Salvador, Uruguay, Suriname, and Guyana.

3 See https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-informal-y-seguridad-social.

4 Chile, Uruguay, Brazil, Argentina, El Salvador, Venezuela, Mexico, Dominican Republic, Guatemala, Colombia, Nicaragua, Ecuador, Bolivia, and Peru.

low educational attainment) are strongly correlated with characteristics of the informal economy such as lack of social protection. These results suggest that criteria based on firm size (in the informal sector) are a suitable approach to calculate informality rates, at least for the Colombian case.

Thus, this document uses the informality definition based on firm size because: 1) the results in Colombia (and in Latin America) suggest that this definition is an appropriate approach to measure informality; 2) the Colombian government adopted this definition as an official statistic to measure informality; and 3) one of the primary purposes of this book is to compare official labour market statistics with vacancy data to test the representativeness of job portals (see Chapter 8).

The magnitude of the informal economy problem depends on different processes. On one side, there is an "exclusion" process. More specifically, workers and companies would prefer formal jobs with benefits mandated by the State; however, some barriers restrict agents' access to the formal economy. These restrictions or barriers can take different forms, such as excessive taxation or lack of certain worker characteristics (e.g. skills), that make it difficult to enter the formal economy. This framework suggests that informal firms and workers are a disadvantaged group.

On the other side, some workers and firms voluntarily choose to remain in the informal economy, based on their preferences for work and the net benefit of being in the informal versus formal economy. In order to belong to the formal economy, workers and firms need to incur certain costs, such as tax revenue, health, and work insurance, and, in return, the state must provide benefits, such as health care, access to credit, etc. However, these benefits might not compensate for the cost of formality (such as taxes). Thus, informal economy can be an "escape" for workers and firms to avoid the formal economy and its failures related to the provision of services (Perry et al. 2007). These facts highlight that the benefits of being in the formal economy are not enough to move some agents into this economy.

Informal economy is usually a term that describes individuals working in unregulated jobs and is associated with inadequate working conditions, lack of social security, lower productivity, limited access to the financial system, etc. As Perry et al. (2007) pointed out, the size of the informal economy is relevant because it affects a country's productivity and growth. Informal firms might experience more barriers to access credit, broaden their sale markets,

and innovate, which might reduce their potential productivity. For instance, the lack of social protection and other work risks might result in a lower incentive for establishments to invest in human capital (see Section 2.4) and lead to lower worker productivity.

Informal economy, along with unemployment, is considered one of the most important indicators to measure well-being in the labour market (OIT 2013; Mondragón-Vélez, Peña, and Wills 2010). Both phenomena are prevalent in Latin American economies and reflect a vast underutilisation of the labour supply. This result reveals the inability of Latin American economies to generate “quality” employment for those who want to work and can work (ILO 2019a). For these reasons, it is essential to measure and consider informal economy in the analysis of any country’s labour market, especially in countries like Colombia where the informality rate is comparatively high, at around 49.4% in 2016 (DANE 2017a) (see Chapter 3).

To conclude this subsection, informal economy is a relevant phenomenon, which affects different socio-economic outcomes, such as productivity, social protections, etc. The high incidence of the informal economy in Latin American countries like Colombia makes it an important factor to be considered in Colombian labour market analysis. However, this term might cover a variety of activities that can be measured in different but correlated ways. Despite some limitations, the Colombian literature suggests that a valid criterion to classify workers as informal is based on company size, which is the one adopted in the official Colombian labour market statistics and in this book.

Related to unemployment, the informal economy phenomenon might arise due to an extended number of factors; rigid wages, comparatively high non-wage costs, technological shocks, and discrimination (e.g. gender preferences) are examples of such factors, and a vast body of theoretical frameworks have been developed to analyse their role. One of these theoretical frameworks stresses the importance of skills in labour market outcomes, such as unemployment and informal economy. Individuals possess different labour characteristics that make them more or less productive for specific jobs (Albrecht, Navarro, and Vroman 2007), so while companies hire labour with different attributes to perform different tasks and produce their products, the misallocation between the skills possessed by workers and the skills demanded by employers might influence unemployment and informality rates.

This framework might be applied in a context such as Colombia where there is a comparatively high portion of companies complaining about the skills of the labour supply, and at the same time there is a high proportion of workers desiring formal jobs (Chapter 3 provides a detailed discussion of the Colombian context). Thus, worker skills are important for the economy, which is examined in more detail in the following subsection.

2.2.4. Skills

Skills are a relevant factor that have strong implications for employment outcomes, such as productivity, wages, job satisfaction, turnover rates, unemployment, informal economy, etc. (Acemoglu and Autor 2011; OECD 2016a). However, the concept of skills can be understood and interpreted from different perspectives: social constructionist, positivist, and ethnomethodological, among others (Attewell 1990; Green 2011; Warhurst et al. 2017). Additionally, there are multiple typologies of skills (e.g. worker skills and skills as attributes of jobs). Thus, this section discusses the definition of skill adopted in this document to analyse labour demand based on information from online job portals.

2.2.4.1. Defining skills

Each school of thinking emphasises the importance of different elements that should be considered by the concept of “skill.” Within the social constructionist school, for instance, skills are a complex construction of job tasks, labour supply and demand, and certain social conditions (Vallas 1990). Consequently, skills are defined by the tasks associated with each job, together with the capacity to enclose a number of people into a profession or career. Therefore, as Gambin et al. (2016) pointed out, from a social constructionist perspective social “norms” and task complexity determine what a valued skill means. This approach is part of an ongoing, subjective, and extended debate in which it is difficult to delimit what social processes might affect the construction of skills in a particular society. Consequently, the social constructionist school often finds it challenging to generalise and compare skills between different societies or groups (Green 2011).

The positivist approach emphasises other aspects. This approach states that skills are objective attributes of individuals or jobs, which are independent of the observer. This view focuses on obtaining uniform skill measures to provide the most precise skills indicators for positivist-based research (Attewell 1990).

Even though there are different ways to define "skills," most perspectives agree that the concept of skills is strongly related to task complexity required to carry out a specific job. In concordance with Green (2011, p. 11): "all skills are social qualities, yet are rooted in real, objective processes, not in perceptions." Thus, this book interprets skills as attributes of people or jobs, which are required to perform certain tasks in the labour market. Consequently, in this document, skill refers to any measurable quality that makes a worker more productive in his/her job, which can be improved through training and development (Green 2011). Simply put, according to Gambin et al. (2016), a skill refers to "the ability to carry out the task that comprises a particular job."

This perspective might be particularly helpful to ease the operationalisation of skills into quantitative measurements (to provide easily measured variables), as well as to enable policymakers and researchers to obtain precise quantitative results to produce straightforward public policy recommendations (Attewell 1990), which is also one of the main objectives of this document. However, this positivist viewpoint has some limitations; for instance, to measure a skill with a variable like years of education could be considered reductionist. As will be discussed in the next subsection, variables like education might fail to properly measure skill acquisition and job performance (Attewell 1990).

Despite the limitations present in all schools of thinking, a positivist perspective (frequently presented in economic studies) is adopted in this document in order to provide imperfect but sufficiently reliable and valid indicators for public policy recommendations regarding skills within vacancy data on online job portals. This definition of "skills" still encompasses many elements, such as qualifications, competences, education, and aptitudes, among others (Green 2011), which can be measured by different indicators depending on the typology used and the tools available to measure those qualities (skills). The economic literature has used a variety of proxies to measure the different dimensions of skills in the labour market, some of which are limited, since a portion of the typologies overlap, while others do not make a clear separation between skill categories (as will be explained in more detail in the next subsection).

Given that multiple typologies of skills are used even within the same economic discipline, it is necessary to discuss which are the most appropriate ones for this book. These different typologies can be organised into two groups: those that focus on the worker's skills and those that use a task-based approach.

2.2.4.2. Worker skills

At an early stage, human capital theory stated that necessary skills for work could be obtained through education (Becker 1962; Mincer 1958). In consequence, educational attainment is regarded as a way to define skills. An educated worker is considered highly skilled and, thus, more productive when he/she accumulates more years of education and experience. Accordingly, increased human capital through education (the main source of scientific knowledge) is thought to increase employee productivity in a range of tasks (Attewell 1990; Becker 1962; Mora and Muro 2008).

Consequently, the accumulation of skills (in terms of knowledge) rather than the use of skills toward specific jobs has been the focus of analysis for academics and policymakers (Becker 1994; Psacharopoulos 1985; 2006). However, the economic literature has found that educational attainment only explains a relatively small fraction of the variance of life accomplishments between individuals (Kautz et al. 2014, p. 9) Additionally, measuring skills by observing educational levels has several limitations. First, educational attainment might be a weak indicator to measure knowledge levels. Education (or qualification) is acquired before people start to participate in the labour market; however, those qualifications might not be appropriate or might depreciate over time, compared to other skills learnt in the workplace.[5]

Second, Becker (1994) acknowledges that educational measures ignore some sources of learning, and Cunha and Heckman (2007) suggest that skill formation/acquisition occurs through a variety of processes and situations. For instance, skills can be acquired outside of schools, through on-the-job

5 For instance, with the emergence of modern devices (e.g. computers), new technologies have been introduced in the labour market to perform different jobs (such as programming, social media manager, etc.), which, in general, were not taught in the educational system years ago. Thus, for some jobs, being up-to-date and being able to use these new technologies can be considered more valuable for the labour market compared to years spent in education.

training (such as apprenticeships, coaching, etc.) and/or off-the-job training (such as lectures, simulations, etc.). An extended literature review on labour economics shows the effects of job training on different outcomes. Bassanini et al. (2007, p. 128) completed an exhaustive review of data resources (Continuing Vocational Training Survey, CVTS; International Adult Literacy Survey, IALS, among others) for on-the-job training in Europe. The authors found evidence that on-the-job training has a positive correlation with private returns for employees and employers (Bassanini et al. 2007, p. 128). Likewise, Asplund (2005), Barrett and O'Connell (1999), and Blundell et al. (1999), among others, extensively reviewed the different effects of off-the-job training on social and private outcomes. Most of the studies reviewed found a positive impact on social and private returns.[6]

Third, education variables do not take into account other skills generated via learning-by-doing in the production process. People continue to learn new skills and reinforce them through repetition (Arrow 1962; Dehnbostel 2002; Rutherford 1992). Different empirical studies show that these learning processes increase a firm's productivity. For instance, Bahk and Gort (1993) observe that in 15 industries in the US, learning-by-doing generates skills (knowledge) and reduces the production costs of incumbent, established organisations.

Finally, employers not only require cognitive and academic skills (qualifications), but also consider personal characteristics as important elements to perform a job. As Green (2011) and Grugulis, Warhurst, and Keep (2004) note, companies have labelled behavioural characteristics (e.g. reliability, responsibility, leadership, motivation, politeness, and commitment, among others) as skills needed in the production process. It is not just the knowledge learnt through formal education, job training or learning-by-doing that produces more-skilled workers; personal characteristics, such as traits, behaviours, and attitudes towards work are also considered as skills (Grugulis, Warhurst, and Keep 2004; Kautz et al. 2014). For instance, Brunello and Schlotter (2011) and Lindqvist and Vestman (2011) note that wages tend to be higher for workers with higher non-cognitive skills, while people with low non-cognitive skills are

6 Nevertheless, some studies suggest that off-the-job training might have greater impacts on productivity than on-the-job training in US manufacturing industries (for example, Black and Lynch 1995).

significantly more likely to become unemployed. In contrast, when Cunningham and Villaseñor (2016) reviewed 27 studies on the skills-demand profiles of employers in developed and developing economies, they found a greater demand for socio-emotional[7] and higher-order cognitive skills[8] than for basic cognitive[9] or technical skills.[10]

Given the importance of the behavioural characteristics of workers and analysing these skills, broader typologies have been recently adapted to measure more of these skill dimensions. For instance, Green (2011) notes that contemporary approaches favour the categorisation of cognitive,[11] physical, and interactive skills.[12, 13]

2.2.4.3. Skills as attributes of jobs

As an alternative to the above worker skills approach, other typologies focus on the attributes of jobs rather than the attributes of a person to measure job complexity. More complex activities require greater skills (Attewell 1990; Green 2011); thus, task-based typologies have become widely used in the economic literature on labour market because these typologies provide a framework to describe processes and changes in job tasks, such as job polarisation,[14] and the effect of implemented new technologies in the occupational structure (Acemoglu and Autor 2011; Autor and Dorn 2012).

---

7 Socio-emotional skills are behaviours, attitudes, and traits that are considered necessary complements to cognitive skills in the production process.

8 Higher-order cognitive skills comprise the capacity to deal with complex information processing. These tasks include critical thinking, application of knowledge, analysis, problem-solving, evaluation, oral and written communication, and adaptive learning.

9 Basic cognitive skills comprise fundamental academic knowledge and comprehension, including literacy and mathematics.

10 Technical skills are defined as the specific knowledge required to carry out an occupation.

11 Cognitive refers to areas where thinking activities, such as reading, numeracy, and IT, among others, are required.

12 Physical skills are task-related, which refers to dexterity and strength, while interactive skills comprise all forms of communication, including emotional and aesthetic behaviour.

13 For a more detailed description of other typologies used to categorize the behavioural characteristics of workers see Green (2011).

14 Job polarisation consists of a decline in the employment share of middle-skilled cognitive and manual jobs characterised by routine tasks.

Occupation classifications appear to be the most common task-approach used in the economic literature. According to the ILO (2012b, p. 59), an occupation can be defined as a "set of jobs whose main task and duties are characterised by a high degree of similarity." Occupational groups or titles are constructed by a group of experts who survey different workplaces and observe workers doing their jobs (ILO 2013).[15]

Nevertheless, this occupational approach has its limitations. Within occupations, skill levels or the kinds of skills being utilised can differ depending on the sector, company size or the country (UKCES 2012). Moreover, occupation classifications are not updated as fast as labour market changes occur. For instance, the International Standard Classification of Occupations (ISCO) is updated approximately every ten years; yet, between these processes and periods, many changes in terms of skills can occur. Thus, prevailing occupation classifications can be obsolete when analysing actual labour market skills.

Another limitation worth considering is that most occupation classifications do not take into account personal features, such as attitudes, traits, and values. An exception can be seen in the Occupational Information Network (O*NET) system in the US, which contains information on hundreds of standardised and occupation-specific descriptors. It describes occupations in terms of knowledge, skills, and abilities required by workers, as well as how the work is performed in relation to tasks, work activities, and other descriptors (onetcenter.org 2016).

Given the above labour market concepts (supply, demand, unemployment, informal economy, and skills, among others), the literature has provided a theoretical framework that helps to understand the labour market dynamics of interest for this document. The following two sections present the main theoretical model for this study in order to explain why skill mismatches might arise, as well as their relevance and implications for labour market outcomes such as unemployment and informality.

---

15 While different occupational classifications exist, like the SOC (Standard Occupational Classification) in the US, every classification system agrees with the ILO's basic definition of occupation. The main differences emerge in the grouping of each occupational category.

2.3. How the labour market works under perfect competition

The third section of this chapter describes the labour market and its main outcomes, such as unemployment, wages, etc., under the assumption of perfect competition. At an early stage, to analyse the matching problem between the demand and supply of skills in the labour market, scholars in the field of economics have developed a basic theoretical framework based on the assumptions of perfect competition (Cahuc, Carcillo, and Zylberberg, 2014). This framework outlines that, on the one hand, employers faced with a need for labour services (a derived demand, based on the demand for their product) create job offers with certain requirements (skills), and, on the other, existing employees and new applicants with those characteristics accept these jobs when the wage offered is more than their reservation wage.[16]

2.3.1. Labour demand

The labour market works under perfect competition when employers and workers are perfectly informed about the quality (e.g. job requirements, localisation of job opportunities, etc.) and price of "labour" (e.g. wages), all agents are price-takers (which means that there are no monopolistic/monopsonistic powers), and there is perfect human rationality (all agents are capable of analysing all possible economic decisions and outcomes, and choosing the path that maximises their utility or profits) (Cahuc, Carcillo, and Zylberberg, 2014; Sen 1977). Given these assumptions, what defines a labour market can be expressed as follows.

On one side, picture a representative firm, which produces goods and services by using two inputs, Labour (L) and Capital (K), at a certain technology level. Consequently, the production function (F) of this representative firm is given by:

---

[16] Cappelli (2015) points out that there is another theoretical framework that explains the relationship between labour supply and employer demand. Employers can select general skills at entry-level positions and train their employees over a working lifetime to develop specific skills needed for the company. However, the same author notes that this approach has become less plausible in recent years because employers tend to hire applicants who already have the specific skills they require.

$$Y = F(L,K)$$

where Y denotes the physical output of the firm and pY is its value-added form, where p is the market price of the product. The cost of labour used in the production takes the form of wages and other on-costs, such as national insurance (the price per hour of hiring a unit of labour services), while the cost of capital is the price of renting a unit of capital.[17]

In the short run, when capital is fixed, the marginal product of labour falls as the number of employed individuals rises. The initial condition for employing anyone at all is that the value of the marginal product of the first worker exceeds their going wage; if so, the firm expands its number of employees until the marginal return to the last unit of that labour equals the marginal wage (cost of labour):

$$p\,F'(L) = w$$

On the other side, there are a large number of workers who offer a certain quantity of labour and will receive a wage if they are hired.

2.3.2. Labour supply

The utility function of a representative worker is composed of two parameters: income (R),[18] which is equal to their wage (times the number of hours worked if the worker is hired, and zero if the worker is not hired),[19] and the individual's leisure time (the number of hours not spent at work). There is decreasing marginal utility of income (which is spent on goods and services or saved) and leisure time; thus, the line in the utility function for combinations of income and leisure to yield a given level of utility (i.e. an indifference curve) is convex to the origin (zero income and leisure). Indifference curves further from the origin are associated with higher levels of utility.

---

17 This is a first approximation because it may be cheaper for the firm to buy capital goods in order to avoid paying profit to those who rent out the goods. The rental price is the rental firm's estimate of forgone interest, plus depreciation of the capital, plus their profit.

18 For simplicity, it is assumed that other forms of income do not exist.

19 It is assumed that the total income of workers is consumed by different goods and services.

2.3.3. Market equilibrium

An equilibrium in the market is achieved where the upward sloping labour supply curve cuts (in other words, it equals) the downward sloping labour demand curve at a certain level of wages (w^*) (labour supply equals demand). Only individuals whose reservation wage (θ) (reflecting their disutility of work) is greater than the equilibrium wage do not participate in the labour market (inactive). In a perfect competition model, as there is perfect information in terms of labour supply and demand, all individuals who wish to participate in the labour market ($\theta \leq w^*$) will find a job, and firms will find a worker to fill their vacancies.

The model does not explicitly talk about the role of skills in the labour market; yet it is relatively easy to incorporate this aspect into this labour market model. As mentioned above in Section 2.2, Mincer (1958) and Becker (1962) introduced the idea that education is an investment in the economic model. Thus, education makes an individual more productive and might create wage differentials.

To be more specific, if people know the relevant characteristics of each job (perfect information), they can choose a general level of training (i) that will increase their production function: y(i). Firms will demand workers with a certain level of training (i) until the marginal benefit of using one unit of that labour equals the marginal wage: w(i). Consequently, the wage of a worker, w(i), will be a function of the level of qualification, all other things being equal, and the possibility of a higher wage acts as an incentive for training. Thus, individuals will train until the marginal cost of training equals the marginal return of this investment. Once more, under the assumptions of perfect competition, an equilibrium is reached when labour demand equals labour supply, and all individuals who wish to participate (θi ≤ wi*) will find a job.

Therefore, one of the most remarkable results from this model is that under perfect competition there is no structural unemployment, instead all workers receive a wage (w*) at their level of employment (E*) (Figure 2.3). Nevertheless, there is a possibility that unforeseen impacts on the supply of labour might create disequilibrium in the short run (Bosworth, Dawkins, and Stromback, 1996, p. 200). For instance, as shown in Figure 2.3, improvements in technologies such as computers might increase the demand for people who know how to use that technology (from D_0 to D_1), and, consequently, wages will rise from

W_0 to W_1. This situation might create a scarcity (E_1*-E_0*) of those people for a period. However, as all agents are (somehow) well informed, workers will start offering labour according to employer requirements.

Figure 2.3. **Labour market equilibrium under perfect competition**

Source: Author's elaboration.

When job seekers know the job requirements (skills, experience, occupational requirements, etc.) and the localisation of job opportunities (cities, companies, etc.), they will train and search in the proper places where vacancies are available. Moreover, as employers know the characteristics of job applicants (e.g. skills), they will hire people who match their job requirements. Additionally, education and training providers (as any other firm) will have all the relevant information to create and adjust their curriculum contents according to employer requirements.[20] Thus, people will find a job according to their characteristics (skills) and employers will find workers according to their requirements. Hence, unemployment rate remains comparatively low under perfect competition and there are no barriers that force a worker to work involuntarily in the informal economy.

20 Note that for education and training providers, to offer the "right" skills, it is necessary to assume that there are no institutional barriers. For example, providers must have the capacity to invest in the equipment necessary to train people in the skills required by employers.

According to the perfect competition model, people make optimal decisions based on the options (information) available to them. Thus, the perfect information assumption is a key element for the well-functioning of the labour market because it helps people to choose the option that maximises their utility or profit. When there are information problems, even fully rational agents in a non-monopolistic labour market might not know the option that could provide the maximum utility/profit. In the labour market, this information problem means, for instance, that job seekers and training centres might not know what skills are being demanded. Some people might acquire the “wrong” skills according to labour demand. Consequently, there will be people with certain skills who are excluded from the formal economy because their skills are not being demanded, and there will be unfilled vacancies because there are no people with the proper skills.

2.4. Market imperfections and segmentation

To develop the above ideas, Section 4 explains how imperfect information (e.g. labour market failures) might increase skill mismatching and, consequently, might create labour market segmentation between formal and informal workers along with a comparatively high unemployment rate.

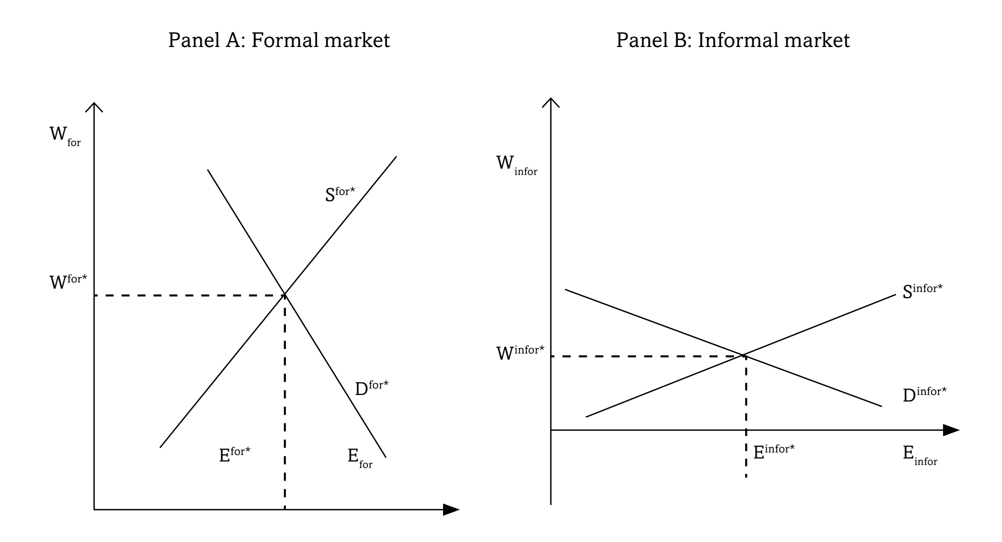
2.4.1. Segmentation

The above assumptions about perfect information (defined in Section 2.3), where all agents in the model are price-takers and rational, are too simplistic (Garibaldi 2006). An extended literature review has shown that the high incidence of informality in countries like Colombia can be due to labour market segmentation (Doeringer and Piore 1971; Reich, Gordon, and Edwards 1973) (see Chapter 3). Specifically, barriers might exclude some workers from the comparatively high-productive sector (e.g. formal sector) and drive those excluded individuals to a more disadvantaged sector, such as the informal market (Piore 1972; Gambin et al. 2016; Palmer 2018).

This duality of the labour market is represented in Figure 2.4. Panel A depicts the more productive formal market sector in which equilibrium wage

is W_{for}* at a level of employment E_{for}*. While Panel B illustrates the more disadvantaged segment in which the equilibrium wage is W_{infor}* at a level of employment E_{infor}*.

Figure 2.4. **Labour market segmentation**



Source: Bosworth, Dawkins, and Stromback (1996, p. 199).

By comparing Panel A and Panel B, two aspects arise. First, labour demand and supply in the formal sector is comparatively more inelastic than the informal sector. This result reflects the fact that in the formal market there are more labour regulations (such as minimum wages, non-wage labour costs, etc.) and more training time, among other entry costs, that make supply and demand less responsive to changes in wages than in the informal sector. Second, wages in the formal sector are higher than in the informal (disadvantaged) sector (W_{for}* > W_{infor}*); consequently, this outcome shows that there are incentives to be part of the formal sector. However, there are some barriers that prevent people from entering the more advantageous segment of the labour market.

The economic literature reveals several barriers that might explain this labour market segmentation (Reich, Gordon, and Edwards 1973). One of these barriers is that potential workers possess imperfect information about the skills required to fulfil employer requirements. Imperfect information might explain

why some workers, even when there are incentives (e.g. higher wages) for them to belong to the formal segment of the labour market, remain outside of this more advantageous sector, while some vacancies remain unfilled.[21] Thus, as $W_{for}^* > W_{infor}^*$ and the labour $conditions_{for} > conditions_{infor}$, there is an incentive for workers to develop the required skills to transfer from the informal to the formal sector, although doing so might take time.

2.4.2. Imperfect market information

As pointed out by Gambin et al. (2016), there are different causes of imbalances (imperfections) in the labour market. For instance, there might be capital constraints, uncertainty about future demand, labour market immobility, institutional barriers, etc., which prevent people from making investments in training or from mobilising workers to places or sectors that require certain skills. However, as previously mentioned, perfect information is one of the necessary conditions for the well-functioning of the labour market (but it is not a sufficient condition). This assumption supposes that all workers know the particularities (e.g. skills required, wages, among others) of all available jobs, and they only need to decide the quantity of labour (number of hours) offered that they are prepared to work, while firms know the characteristics of all potential workers and can choose the one who most suits their job requirements, and education and training institutions offer programmes that are aligned with employer needs. However, labour market failures arise due to imperfect information, which occurs when agents in the economy (in this case, employers, employees, and training centres) are not fully informed about the price or quality of the product they are going to buy or sell. Therefore, agents might not make optimal decisions (Stiglitz et al. 2013).

For instance, education and training institutions need to have up-to-date labour market information (e.g. skills and occupational requirements, number of people demanded, etc.) to design (curriculum contents, number of courses, etc.) and offer programmes that cover the needs of the labour market. However, training centres (usually) do not have the necessary means and resources to

---

21 As will be shown in more detail in Chapter 3, evidence suggests that this situation is prevalent in countries like Colombia.

know employer requirements (see Chapter 3 and 4). Given the difficulties in obtaining proper labour market information, education and training providers cannot respond properly to labour market changes. As mentioned by Almeida, Behrman, and Robalino (2012), this lack of proper information prevents education and training programmes from being aligned with labour demand needs. Consequently, misaligned, outdated or low-quality curriculum contents will arise due to imperfect labour market information (see Chapter 3). People might not have the "right" skills, and companies might not find workers with the skill sets they need. Thus, limitation on information might create phenomena such as skill mismatches. In particular, skill mismatches occur when there is imperfect information in the job search process or in the workplace regarding the particularities of jobs; mismatches that misalign the labour demand and supply of skills (UKCES 2014). These phenomena can acquire different forms, such as skill gaps, skill surpluses, and skill shortages, with various consequences on the economy, such as unemployment, informality, job dissatisfaction, among others.

Once a job match is completed, employers might realise that their current employees need more skills to be completely proficient in their jobs; this problem is called a skill gap and is considered part of the phenomenon of skill mismatch.[22] Nevertheless, the definition of skill gaps per se does not capture the entire skill mismatch phenomenon. For instance, a skill surplus might occur within workplaces. This term refers to a situation where a certain job does not require the highest level of an employee's competences (Adalet McGowan and Andrews 2015). According to Green and Zhu (2008), graduate over-qualification (which is a way to measure skill surpluses) was about 33% in the UK in 2006. This underutilisation of labour supply creates a misallocation of education and training resources (money and time are invested in programmes not demanded by the labour market), as well as increases job dissatisfaction (people do not fully use the skills they possess: underemployment) and employee turnover, which might be due to a loss of pay from being over-qualified (Green and Zhu 2008; Okay-Somerville and Scholarios 2013).

22 Several economic studies have shown the importance of skill gaps in the economy. For instance, in an Irish-based study, McGuinness and Ortiz (2016, p. 19) suggest that the phenomenon of skill mismatch increases labour costs by approximately 25%, thus negatively affecting the competitiveness of Irish firms.

However, given the multiple configurations that the skill mismatch problem encompasses and labour market data available to analyse an economy such as Colombia's, hereinafter this study will focus on skill shortages. This term refers to issues that arise in the job searching process when there are no applicants, or the applicants do not have the minimum level of skills needed to carry out the tasks required by employers. There is a skill shortage when the labour supply lacks skills in relation to what employers currently demand to fill their vacancies (Green, Machin, and Wilkinson, 1998).[23, 24]

Claims of skill shortages have been made globally. For instance, the European Company Survey for Spring 2013 reports that around 39% of firms in Europe experienced difficulties in finding workers according to skill requirements (Cedefop 2015, p. 20). Similarly, the ManpowerGroup (a well-known international consulting firm) carries out a Talent Shortage Survey, where employers around the world are asked whether they have difficulties in filling their jobs (Mazza 2017). As reported in 2016, due to skill shortages, 40% of the companies interviewed worldwide faced difficulties to fill their vacancies (ManpowerGroup 2016). However, in countries like Colombia this phenomenon is even larger (as will be shown in more detail in Chapter 3).

The human capital framework in economics has developed different theories to consider the possibility of imperfect information, as well as to explain labour market outcomes in a more realistic way. The search and matching theory, for example, has become one of the most prominent theories to explain skill mismatches and their relation to unemployment (Andrews et al. 2008). This model states that vacancies and workers are heterogeneous in terms of one characteristic: skills. However, obtaining information about the price and quality of labour can be costly, and not everyone has access to this information, which is a limitation that might affect the behaviour of workers and firms.

With imperfect information, the opportunity cost (θ parameter) is not the only relevant parameter to determine whether a person is employed or not. In addition, individuals need to devote time to find a job and firms might need to wait or search actively for the candidate that suits their requirements. Thus,

23 This definition excludes other causes of shortages such as firm size and lack of union recognition, among other causes (Green, Machin, and Wilkinson, 1998).

24 Chapter 9 discusses different possible ways to measure skill shortages.

included here is the possibility that the labour market does not instantly correct mismatches such as skill shortages (hereinafter skill mismatches refer to skill shortages). The efficiency at which the market makes matches between vacancies and workers depends on the matching function (the formation of new relationships such as job formation), which can be expressed as follows (Mortensen and Pissarides 1994):

$$m = m(u,v)$$

Where v represents the number of vacancies, u represents unemployed workers, and m indicates the rate of job matching (number of hired people and vacancies filled) over a given time period. Moreover, m is assumed to be homogeneous of degree one, which means that if u and v are doubled, the number of matches (m) will increase by the same proportion.

Equation 1 can derive the probability that a vacancy is filled:

$$q = \frac{m(v,u)}{v} \tag{1}$$

As vacancies are filled at the Poisson rate, Equation 2 can be expressed as follows:

$$\frac{m(v,u)}{v} = m\left(\frac{u}{v}, 1\right) \equiv q(\alpha) \tag{2}$$

Where α is v/u, and it is interpreted as labour market tightness, an indicator to identify possible difficulties to fill vacancies, or whether it takes a relatively long time to fill an available job.

Employees also make decisions about educational (skills) investments and where to look for a job according to available information. Subsequently, job opportunities reach jobseekers with a certain probability given by the following:

$$p = \frac{m(v,u)}{u} = \frac{v}{u} m \left(\frac{u}{v}, 1\right) \tag{3}$$

Thus, the probability that a worker finds a job and a vacancy is filled is a function of market tightness, which depends on the quality of labour (skills) offered and demanded, among other characteristics.

Vacancies are offered in different places, such as newspapers or online job portals, and the information available there might restrict the number of job advertisements a person screens to make decisions about which roles to apply for. Also, individuals might not have access to or use certain sources that display vacancy information. Consequently, workers' decisions can be based on imperfect information, hence they might or might not properly anticipate an employer's requirements to fill certain vacancies (Mortensen 1970).

Therefore, according to employer requirements, a lack of proper skills (e.g. cognitive and non-cognitive skills) might affect the labour market matching function and create labour market segmentation. Indeed, as mentioned in Section 2.2.4, both cognitive and non-cognitive skills are relevant for the well-functioning of the labour market. As remarked by Desjardins and Rubenson (2011), cognitive skills such as literacy are becoming more important in today's economy due to skill-biased technical changes (e.g. information and communication technologies or ICTs) Moreover, Brunello and Schlotter (2011) and Lindqvist and Vestman (2011) pointed out that people with low non-cognitive skills are significantly more likely to become unemployed. Thus, the combination of cognitive and non-cognitive skills demanded by employers and possessed by job seekers will considerably determine the performance of the matching function and other labour market outcomes, such as unemployment and informality rates.

If the likelihood of finding a formal job is relatively low (which might mean that companies do not demand the cognitive and non-cognitive skills some workers have attained), it can take time to find a job. Individuals whose skills are not in demand in the labour market have two options: 1) they can continue searching or create a job for themselves through self-employment, or 2) take an informal job as a way to earn an income and fulfil personal and family responsibilities. Those individuals who value an informal job more than the expected value of searching for and taking a job in the formal sector will be part of the informal economy (Albrecht, Navarro, and Vroman 2007). From the other perspective of the labour market, firms might not be able to gather perfect information about the skills possessed by potential individuals and have knowledge about where to find them (Desjardins and Rubenson 2011; Oyer and Schaefer 2010). According to this view, employers will hire an

individual when the expected value of matching that individual exceeds the cost of posting a vacancy[25,26] (Burdett and Smith 2002).

Consequently, hiring is an important and costly selection process for heterogeneous productive individuals and firms, and its efficiency depends on the research behaviour of employers and job searchers, as well as on the information available to them (Banfi and Villena-Roldán 2019). In this sense, companies can face some difficulties in finding people that meet their requirements. Due to that, they spend significant resources on advertising, posting job vacancies, and screening to select appropriate workers (Autor 2001).

Even with those strategies in place, it is possible to reach a situation where unemployed or informal workers with certain characteristics (skills) are willing to work in formal jobs and vacancies available to be filled. This situation can occur because the skills possessed by job seekers are not those required by the companies, resulting in skill shortages (or a skill mismatch).

Provided that companies require different skill combinations, and workers have restricted access and limited capacity to respond to those requirements, one straightforward solution to tackle this phenomenon and its consequences is to lower the cost of having (relevant) information about the current labour demand for skills. By doing so, workers have proper insights about current job roles, which might shape their decisions to acquire skills according to employer requirements. The matching function will become more efficient if workers have less imperfect information about the employers' needs, and thus unemployment and (involuntary) informality will be reduced.

Moreover, the role of education and vocational education and training (VET) systems is relevant to reduce skill mismatches. Education and VET systems are one of the main ways to prepare (deliver skills to) people for work

25 When the cost of posting a vacancy exceeds the profit to be gained from the match, employers do not post vacancies (Burdett and Smith 2002).

26 Other models also recognise that employers might not possess perfect information about worker skills. For instance, Spence (1978) developed a job-market signalling model where employers are not sure about the "productive capabilities" (skills) of a potential employee. To overcome this issue, employers believe that credentials, such as higher education, are positively correlated with a worker's "productive capabilities." Consequently, potential employees need to send a signal about their skill levels to potential employers by acquiring credentials. In this case, credentials are considered as a proxy to measure skills and help employers and employees in the hiring process.

(Green 2011; OECD 2014a), and they might be affected by a restricted access and a limited capacity to analyse and anticipate employer requirements. Consequently, it is almost pointless that workers have the right information about current employer requirements for skills, that is, in case if there are no education and training systems in place that provide them. In consequence, the better understood how to adopt and develop this knowledge into education and training programs and into worker decisions, the better the match will be between worker skills and vacancies (Cedefop 2012a) (see Chapters 9 and 10).

2.5. Conclusion

This chapter has outlined a basic labour market framework in order to properly use vacancy data and address the phenomena of unemployment and informal economy. The labour market is a space where workers (labour supply) offer a quantity of "labour services" with certain qualities to fill vacancies, and employers (labour demand) hire this merchandise at a certain price (wages). In terms of the labour market, people can be divided into three groups: 1) workers whose labour services are bought by employers in the formal economy; 2) workers employed in the informal economy, who are characterised by a lack of social security, limited access to the financial system, etc.; and 3) workers who offer their labour services but are not hired by employers (unemployed). The size of each group depends on different elements. However, the literature discussed in this chapter stresses that skills are a relevant factor to determine labour outcomes, such as unemployment and the size of the informal economy.

Due to their importance and multiple dimensions (e.g. qualifications, competences, education, aptitudes, etc.), skills can be defined in different ways; nevertheless, most of those definitions link the task complexity of each job to the characteristics that each worker needs in order to successfully carry out these job tasks. For this reason, this book considers that a skill is any measurable quality that increases worker productivity and can be improved by training and/or development. Using this definition, it is possible to analyse and extract information on vacancies to construct more reliable indicators of the skill level required by employers (e.g. qualifications), as well as to address possible skill mismatch issues.

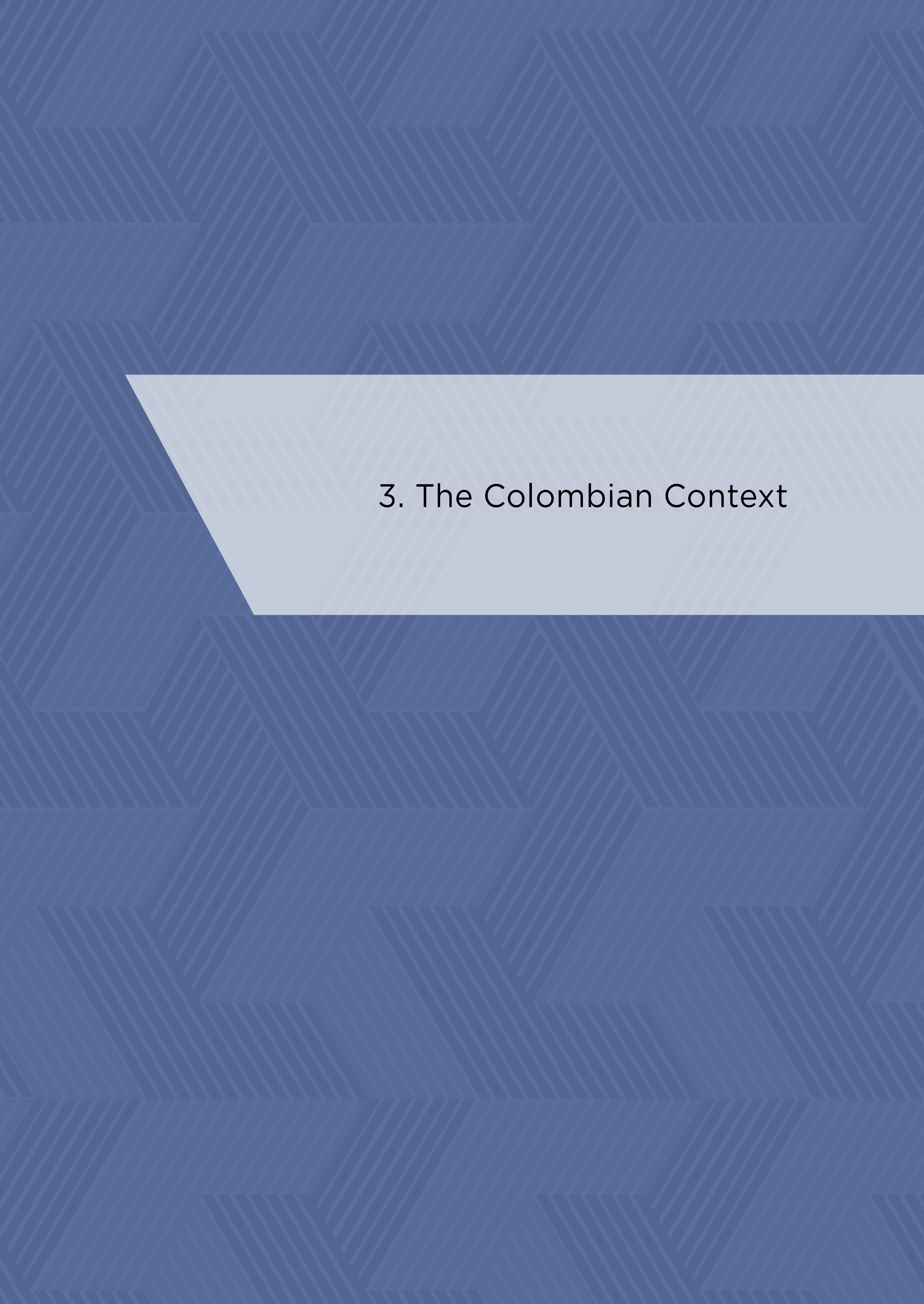
Under perfect competition, the over- or undersupply of skills (skill mismatches) only arise in the short term and have relatively small implications for unemployment and informality rates (exclusion). However, the conditions required for perfect competition rarely exist because agents have imperfect information about offered and demanded skills. This imperfection in the labour market might create a situation where there is a lack of skills in relation to what employers currently require in order to fill their vacancies: a skill shortage. Skill shortages might create labour market segmentation where workers with the "right" skills have more probabilities to belong to the formal economy, while workers without the "right" skills (according to demand) have more chances of being in the informal economy or unemployed. Consequently, unemployment and informal economy might increase and/or persist over time.

The above skill mismatch problem involves the coordinated actions of at least three different agents in the economy: employers, workers, and education and VET systems. The level of coordination between these three groups determines the extent of skill mismatch. This coordination depends on the availability of information about skill requirements and the capability of workers to process and adopt that information into their decisions, as well as the availability of education and training systems.

In this sense, one way to tackle the skill mismatch phenomenon is to gather information about labour demand, and to extract meaningful information in order to address the decisions of workers and education and VET systems, according to different company requirements. New technological developments offer new opportunities in this respect. This particular theoretical framework and straightforward solution might be especially useful for countries like Colombia where: 1) informality and unemployment rates are high, 2) complaints about skill shortages (skill mismatch) are relatively high, 3) information about company requirements is available from resources such as job portals, and 4) education and VET institutions have difficulties to adapt their programs according to labour demand.

For these reasons, the next two chapters demonstrate that in the context of the Colombian economy, novel sources of information and data analysis regarding the labour demand for skills might have an important effect on public policy and could reduce unemployment and informal economy at a lower cost in terms of time and monetary sources.





3. The Colombian Context

3.1. Introduction

Skill mismatches are a widespread phenomenon that have strong implications on unemployment and informality rates, among other variables (McGuinness and Pouliakas 2016) (see Chapter 2). Nevertheless, some countries display a higher incidence of these issues, which might have severe effects on local labour outcomes. This chapter presents evidence that Colombia is a country where the degree of skill mismatches (skill shortages), unemployment, and informality is relatively high. However, public policies that tackle those outcomes are limited, and, consequently, this makes Colombia a relevant case of study to develop novel ways to analyse and reduce skill mismatches.

Based on the concepts discussed in Chapter 2, this chapter, first, provides an overview of the main characteristics of the Colombian labour market and its evolution over time. Second, it shows that the issue of skill mismatches and their possible incidences have a relatively high impact on the national economy, which needs to be addressed by public policies. Subsequently, it explains the importance of maintaining systems with accurate labour market information to address these phenomena. Finally, it is argued that the lack of information about skill requirements, together with an institutional disarticulation, especially in Colombia (and other developing countries), makes it difficult to develop well-orientated public employment policies to deal with the skill shortage phenomenon. For this reason, there is a need to find novel solutions to systematically provide accurate information and analyse employer requirements and possible skill mismatches.

3.2. The characteristics of the Colombian labour market

This section describes the main characteristics of the Colombian workforce and labour demand in order to present the structure of one of the most relevant labour market issues Colombia has been facing: unemployment and informality.

3.2.1. Labour supply

Figure 3.1 shows the structure of labour supply in Colombia at a macroeconomic level. In 2016, the Colombian working-age population was composed of around 37,851,500 people, of which 64.4% participate in the labour market (approximately 24,405,000 people), representing the current labour supply in Colombia. As mentioned in Chapter 2, labour supply is composed of: 1) people in the working-age population who do not have a job but are looking for one (unemployed), and 2) people who are in the working-age population and are hired by employers (employed), and who are self-employed. As shown in Figure 3.1, around 90.7% of the economically active population (EAP) have a job (employed); however, 5,776,750 people work in informal jobs. In addition, around 9.2% of the Colombian workforce is unemployed.

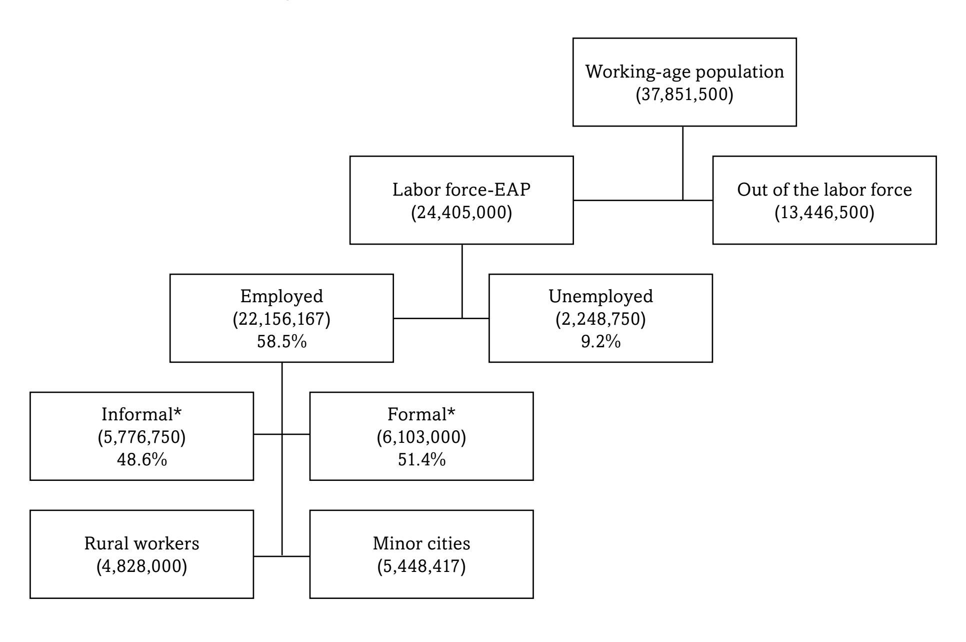
These indicators highlight a key point: in Colombia, the labour participation rate is relatively high. Indeed, it is 2.6 percentage points above the Latin American average (ILO 2016b, p. 29). However, only 51.4% of the employed population has a formal job (Figure 3.1). Moreover, high unemployment and informality rates are persistent in Colombia. As shown in Figure 3.2, in 2001,[27] the annual national unemployment rate was approximately 15%, while the participation rate was 62.4%. In the same period, the informality rate decreased from 50.4% in 2006[28] to 47.5% in 2016 (DANE 2017a). This result means that during the last fifteen years more people have participated in the Colombian labour market. Formal labour demand has absorbed a considerable proportion of labour supply to the point that unemployment and informality rates have declined, even with more people participating in the labour market.

---

[27] In 2001, there were changes in the household survey methodology, which affect the comparison of labour market indicators before 2001.

[28] Due to methodological changes, informality rates are not comparable before 2007.

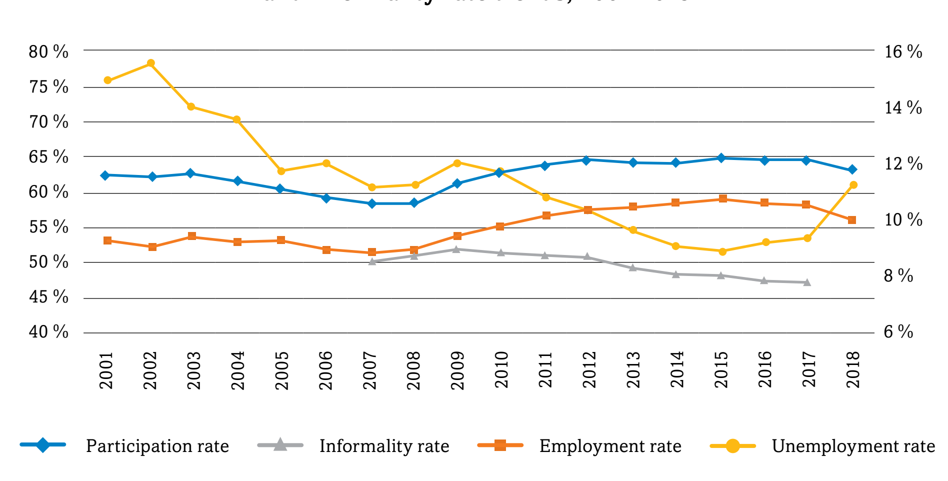
Figure 3.1. **Labour structure in Colombia**



Source: Author's calculations based on DANE 2017a.

*Informality is only calculated for urban areas. As explained in Chapter 2, for rural areas, the definition of informality (company size) is not accurate. At the time this chapter was written, there was no official measure of informality for those rural areas.

Figure 3.2. **Participation, employment, unemployment, and informality rate trends, 2001-2018**



Source: DANE 2017a.

*Unemployment rates are graphed on the right-hand scale.

However, in Colombia, it took a relatively long period (15 years) to decrease unemployment and informality rates to 5.8 and 2.9 percentages points, respectively. Additionally, informality and unemployment trends changed in 2017 and 2018, when the unemployment rate increased by 0.2 and 0.3 percentage points, respectively, and informality rates stagnated around 47%. Although these rates have declined in recent decades, Colombia's unemployment and informality rates are above the world average, and even above the Latin American average (World Bank, n.d.). In particular, in 2015, Colombia was the second economy in the Latin American region with the highest unemployment rate (only surpassed by Brazil), and its informality rate was around 1.4 percentage points higher than the regional average (ILO 2016b). However, informality and unemployment do not affect all workers equally. Table 3.1 shows the general characteristics of the Colombian workforce between 2016 and 2018.

Table 3.1. **Characteristics of the Colombian workforce**

Variables	Formal workers	Informal workers	Unemployed
% General characteristics			
Male	56.7%	53.9%	44.3%
Less than 29 years old	30.5%	23.3%	49.1%
Between 29 and 58 years old	64.9%	62.3%	45.7%
More than 58 years old	4.6%	14.4%	5.2%
% Educational level			
Less than high school	7.0%	29.1%	14.5%
High school	42.5%	53.4%	53.3%
Lower and higher vocational education	21.3%	11.3%	18.4%
Graduate	19.5%	5.2%	11.4%
Postgraduate	9.8%	1.0%	2.4%
Labour market outcomes			
Mean wage (Colombian pesos)	1,511,246	910,508	-
Mean hours worked per week	47.2	43.8	-
Underemployment	31.7%	35.6%	-
Agriculture, hunting, and forestry	2.5%	5.2%	3.4%
Mining and quarrying	1.0%	0.2%	1.0%

Variables	Formal workers	Informal workers	Unemployed
Manufacturing	16.0%	11.0%	11.3%
Electricity, gas, and water supply	1.3%	0.0%	0.5%
Construction	5.3%	8.4%	11.3%
Wholesale and retail trade, hotels, and restaurants	18.9%	42.1%	30.0%
Transport, storage, and communications	6.7%	11.5%	6.8%
Financial intermediation	3.3%	0.4%	1.6%
Real estate, renting, and business activities	13.1%	6.7%	9.2%
Community, social, and personal service activities	31.9%	14.5%	24.9%
Duration of unemployment (weeks)	-	-	20.2

Source: Author's calculations based on GEIH information.

According to the first column, 56.7% of formal workers are male, while the second column indicates that 53.9% of informal workers are male. This result is because in the Colombian labour market more men are working than women. However, the presence of women in the informal market is 2.8 percentage points higher than women in the formal market. Moreover, the third column shows that 55.7% of unemployed individuals are women. These results suggest that unemployment and informality issues are comparatively higher for women than for men.

According to the age distribution of all workers (males and females combined),[29] 30.5% of formal workers were less than 29 years old, compared to 23.3% of informal workers. In contrast, only 4.6% of formal workers were over the age of 58 years, compared to 14.4% of informal workers. However, almost half (49.1%) of the Colombian unemployed population were less than 29 years old, followed by people between 29 and 58 years old (45.7%), and over 58 years old (5.2%). Consequently, older Colombian workers tend to be more exposed to informality, while young workers are more likely to experience unemployment issues.

29 The age distribution presented in Table 3.1 follows the age bands indicated by the DANE, which define a person as young if she/he is less than 29 years old.

The educational distribution[30] shows that the higher the level of education (lower and higher vocational education, graduate or postgraduate), the higher the proportion of formal workers when compared to the proportion of informal workers. Moreover, more than half of the unemployed individuals in Colombia have just a high school certificate. In fact, most formal and informal workers and those who are unemployed only have a high school certificate (42.5%, 53.4%, and 53.3%, respectively).

The monthly average wage of a formal worker is around 1,511,246 pesos (around £377), while the average salary of an informal worker is about 910,508 pesos (around £227). In accordance with Mondragón-Vélez, Peña, and Wills (2010), a formal worker earns 1.6 times more than an informal worker. In contrast, an informally employed person works 3.4 hours less per week than a formal worker. More than one-third of workers are underemployed because of the underutilisation of their skills (skill surpluses; see Chapter 2). However, this percentage is higher for informal workers.

Around 31.9% of formal workers are employed in companies related to community, social, and personal service activities, followed by the wholesale and retail trade, hotels, and restaurants (18.9%), as well as manufacturing (16.0%). In contrast, most informal workers are in the wholesale and retail trade, hotels and restaurants sector (42.1%), followed by community, social, and personal service activities (14.5%), as well as transport, storage, and communications (11.5%). Additionally, most unemployed individuals used to work in the wholesale and retail trade, hotels and restaurants sector (30.0%), community, social, and personal service activities (24.9%), construction (11.3%),

---

30 The general overview of the structure of the Colombian educational system is the following: Pre-school education is for children under six years old, and basic (and compulsory) education is composed of the elementary and middle schools (6th to 9th grades). To have access to higher educational programs it is necessary to have finished high school (10th and 11th grades). People with high school education can choose between lower, higher vocational or undergraduate programs. Frequently, it is not compulsory to have a lower vocational education qualification to access higher vocational programs. When people finish their undergraduate studies, they can continue studying in a specialisation or a master's program. On the one hand, specialisations are programs that usually involve one year of part-time study, in which people can develop and deepen specific qualifications for a specific occupation, discipline, etc. (Ministerio de Educación Nacional 2006). On the other hand, master's programs usually involve two years of full-time study. To be admitted to a PhD programme (in most cases), it is necessary to first obtain a master's degree (OEI, n.d.).

and manufacturing (11.3%). Therefore, the sectors that concentrate most of the informal and unemployed people are the wholesale and retail trade, hotels and restaurants sector, and companies related to community, social, and personal service activities. The last row of Table 3.1 shows that the average duration of unemployment was around 4.7 months (20.2 weeks); this indicates that the duration of unemployment in Colombia is above average compared to the average of the OECD countries, which was 3.6 months between 2016 and 2017 (UK Data Service, n.d.).

The results from Figure 3.1, Figure 3.2, and Table 3.1 confirm that informality is a widespread and persistent problem in the Colombian economy. However, these outcomes can be explained by two different phenomena with different implications for public policy and economic research. As pointed out in Chapter 2, informality might be explained by "exclusion" and "exit" processes. The first term, "exclusion," refers to the situation where there are labour market segmentation and barriers that prevent informal workers from taking formal jobs (with state-mandated benefits). The second term, "exit," occurs when workers and firms decide to stay outside of formality given that the cost of being formal exceeds the benefits of belonging to this sector.

Even though both views (exclusion and exit) are important in Colombia, evidence suggests that exclusion mechanisms are more relevant for the Colombian context. According to Perry et al. (2007), the fraction of informal and independent workers who would rather be formal employees is around 40% in Argentina, 59% in Colombia, and 25% in Bolivia and the Dominican Republic. When informal self-employed workers were asked about their motivations/reasons for being in their current job as an independent worker (such as autonomy, flexible hours, could not find a salaried job, higher wages), their main response to why working as informal and self-employed was that they could not find a salaried job: 59% in Argentina and 55% in Colombia gave this response (Perry et al. 2007, p. 66). Additionally, the authors found similar results for informal salaried workers; thus, difficulties in finding a formal salaried job constitute a much higher fraction of the reported reasons for being in informal salaried jobs than other possible responses.

In consequence, evidence in Latin America shows that a considerable proportion of informal workers would prefer to work in a formal job but cannot find one. Furthermore, the majority of the Colombian unemployed population

(36%) reported in 2016 that the scarcity of available jobs, according to their occupation, is the main reason why they stop looking for formal employment.

This evidence reveals a number of relevant facts: 1) informality and unemployment are relatively high in Colombia, even compared to the country's regional counterparts; 2) labour supply trends reveal that both informality and unemployment rates are explained by structural rather than a cyclical component; that is, there is a significant and persistent portion of people who are looking for a job, however, they are not hired by the Colombian labour demand; 3) most people affected by the phenomena of informality and unemployment are the following groups: less than 29 years old, more than 58 years old, women, and characterised by a low level of education; and 4) a significant share of the workforce employed in informal jobs desires to work as formal workers.

3.2.2. Labour demand

As discussed in Chapter 2, to understand the potential causes of the informality and unemployment results, it is important to analyse also the Colombian labour demand. With a GDP per capita of 14,181.406 US dollars in 2016 (World Bank 2019) (three times less than the OECD average), Colombia is an economy in which employment is high in the service sector. Indeed, this sector encompassed 57.4% of Colombia's GDP in 2013 and employed around 63% of the labour workforce in 2016 (as mentioned in Subsection 3.2.1). Moreover, most employment is offered by micro-, small or medium-sized enterprises. According to the ILO (2014), micro-enterprises (defined as units with up to 10 employees) account for 96% of the country's companies, small enterprises (defined as units between 11-50 employees) represent 3%, while medium-sized (between 51 to 200 employees) and large enterprises (>200 employees) represent 0.5% and 0.1%, respectively. Consequently, 80.8% of the Colombian workforce is employed by micro-enterprises and SMEs (small and medium-sized enterprises[31]), and these enterprises contribute to approximately 40% of Colombia's GDP (OECD 2017a). However, around 60% of those micro-enterprises were in the informal

[31] According to OECD measures, SMEs refer to companies with fewer than 50 employees, while micro-enterprises have, at most, 10 employees or, in some cases, 5 employees (OECD Statistics Portal, n.d.).

sector in 2010 (ILO 2014). All these indicators reveal that there is an important informal economy in Colombia that employs a high number of people in the service sector, specifically, in activities related to sales and retail.[32]

Many factors might explain why labour demand does not fully utilise the Colombian labour force. For instance, the high cost of hiring is one of the main factors that prevent formal companies from hiring more personnel (Bell 1997; Kugler and Kugler 2009; Mondragón-Vélez, Peña, and Wills 2010). Mondragón-Vélez, Peña, and Wills (2010) observe that in the Colombian labour market there are comparatively high non-wage costs (payroll taxes, health and pension contribution, among others) and a high minimum wage relative to the productivity level. These labour market rigidities restrict the formal sector to adapt to the business cycle, thus the size of the informal sector and unemployment increases.

Despite the high cost of hiring in Colombia, there is a relatively high vacancy rate. According to the Human Capital Formation Survey (EFCH, for its acronym in Spanish) carried out by the DANE in 2012 (DANE 2018b), around 80.4% of open vacancies were related to sales and retail activities, and 87.6% and 94.4% to the service and industrial sectors (excluding sales and retail activities), respectively. Moreover, most new vacancies related to sales and retail activities were generated in the area of marketing and sales (68.6%), while in the industrial and services sectors (excluding sales and retail activities), most new vacancies were generated in the production area (66.9% and 82.2%, respectively).

Thus, Colombia's labour demand suggests that, even with the relatively high cost of hiring, while there are formal vacancies available, there are also a high number of unemployed and informal individuals who are willing to work in formal jobs, but who do not do so. Consequently, there is a mismatch between supply and labour demand.

---

32 The DANE carries out a specific survey annually to measure the economic activity of companies related to sales and retail because they possess such a high level of importance in the Colombian market.

3.3. Skill mismatches in Colombia

As presented in Chapter 2, skill mismatches occur where the labour demand and supply of skills are not aligned (UKCES 2014). This misallocation of skills might explain why some countries face high unemployment and informality rates, and, at the same time, a relatively high portion of companies complain about the scarcity of accurate human resources. Consequently, the skill mismatches framework might explain a considerable portion of the labour market outcomes in Colombia (as presented in the previous section).

Globally, Latin America possesses the largest gap between the labour demand and supply of skills (OECD 2017b). In this region, around 44% of companies in 2016 experienced difficulties finding accurately trained candidates (skill shortages) (ManpowerGroup 2016). For Colombia, this rate is even worse, as around 50% of companies face problems filling vacancies due to a shortage of skills (OECD 2017b).

The Beveridge curve for Colombia (that depicts the relationship between unemployment and vacancies to determine how well, or not, job vacancies correspond to unemployed workers) illustrates a deep and constant labour market mismatch (Blanchard and Diamond 1989). According to Álvarez and Hofstetter (2014), Colombia has a relatively high level of vacancies and unemployment, which suggests that a lack of skills in the workforce (skill shortages) is one of the main reasons for Colombia's labour market mismatches.

Moreover, the 2012 EFCH shows that around 62.1%, 67.2%, and 61.7% of employers in the industrial and service sectors, as well as in sales and retail activities, respectively, cited skill shortages[33] as the leading cause of difficulties to find suitable workers. In addition, low productivity/poor performance and lack of specific competences were selected as main reasons to fire workers (around 34.4%, 40.9%, and 33.1% in the industrial and service sectors, and in sales and retail activities, respectively). Thus, the lack of worker skills is a key problem in Colombia, especially in the service sector. In particular, there is a

[33] Some of the categories were: sub-qualified, over-qualified, low performing, gave a bad impression during the interview, lack of candidate experience, lack of reliable information about qualifications and experiences, the candidates did not speak other languages.

large shortage of technical specialists, and a surplus of unskilled workers and middle management professionals (OECD 2015a).

Although the average year of educational attainment has increased to around six years during the last four decades for all age ranges (World Bank n.d.), Colombia remains a country with relatively low levels of education: in 2012, only 42% of Colombian people between 25 and 64 years old reached at least upper secondary school education, around 33 percentage points below the OECD average and just above Mexico in Latin America; whereas only 20% of adults completed a tertiary level of education (12 percentage points below the OECD average) (OECD 2014b). In addition, the Programme for International Student Assessment (PISA), which evaluates education systems worldwide by testing the skills and knowledge of 15-year-old students, reveals a low student performance in mathematics in Colombia. Almost 75% of students fail to achieve the baseline level of knowledge in mathematics, which contrasts with the OECD average of 23%. A low proportion of students (around 0.3%) are top performers, 12 percentage points below the OECD (OECD 2014b). Moreover, based on the Colombian household survey, the Gran Encuesta Integrada de Hogares (GEIH), only 9% of the working-age population during 2014 took a technical or vocational education and training course.

It is not only companies that have observed a large deficiency of skills. Arango and Hamann (2013) consulted an important group of labour market analysts (15 experts) in Colombia about the leading causes of unemployment. The majority (67%) agreed that skill mismatch between labour demand and supply was the main cause of unemployment in the country. Consequently, 60% of the experts recommended strengthening information systems to improve the efficiency of matches between employers and employees.

Thus, there is a generalised consensus between labour market experts and national and international institutions that lack of skills is one of the main reasons for skill mismatches in Colombia. Consequently, as explained in Chapter 2, one of the main issues faced by Colombia is that people, education and training providers, and the government are making decisions about human capital investments based on the currently available labour market information. However, these agents are not accurately anticipating employer requirements to fill their vacancies. Those workers whose skills are not in demand might choose between remaining outside of the labour market (being inactive),

being unemployed or getting employed in the informal sector. Based on the Colombian evidence (discussed above), a high proportion of people select the last two options: the informal sector or unemployment.

At the same time, a relatively high proportion of companies in Colombia complain about the scarcity of workforce according to their needs, which leads to a situation where there are vacancies to be filled. However, due to skill mismatches, the Colombian labour supply does not have the necessary characteristics to fill these vacancies (see Chapter 2). As a consequence, to reduce unemployment and informality problems, information asymmetries between supply (individuals) and demand (employers) for labour must be addressed. Tackling these problems might have a large positive impact on regions like Colombia where unemployment and informality rates are relatively high, and there is a large gap between the labour demand and supply of skills.

As the OECD (2017b) has pointed out, to tackle informality and improve economic stability Latin American countries like Colombia should invest in human capital. The same organisation argues that more education in terms of quantity and quality increases a person's likelihood of finding a job and reduces the probabilities of being unemployed or working in the informal sector. Moreover, to guarantee the effectiveness of human capital investments and to avoid any labour market mismatches as described in Chapter 2 (e.g. overeducation), governments and other institutions need to promote skills that meet company requirements (Gambin, Green, and Hogarth 2009; OECD 2012).

Given the importance of skill mismatches, institutions such as the World Bank (2010), the OECD (2016a), and the ILO (2017b) agree that fostering education and suitable skills (to strengthen human capital) might have a large positive impact on the main employment problems of Latin America (e.g. Colombia). Thus, it is essential for Colombia to achieve at least the minimum skill levels in its population, and to improve the relevance of education and training systems in order to reduce unemployment and promote well-being (OECD 2015a; González Espitia and Mora Rodríguez 2011).

As González-Velosa and Rosas-Shady (2016) mentioned, advanced education and training systems achieve the above by encompassing tools to identify current and future skill requirements for the productive sector. With these tools, curriculum contents can be updated, and the relevance of education and training increased. Consequently, approaches that identify possible skill

mismatches, when combined with a functional system of active labour market policies, can ensure better matches between employers and workers (ILO 2016a).

3.4. An international example of skill mismatch measures

Examples of the above can be found in different regions. As Mavromaras et al. (2013) highlight, the most developed approaches to measure skill mismatches (skill shortages) can be found in the UK. For instance, the Migration Advisory Committee (MAC) built 12 indicators[34] of shortage using data for labour demand and supply. With this set of indicators, the MAC advises to the UK Government on where skill shortages can be filled by immigration from outside the European Economic Area (EEA). In addition, the UK Commission for Employment and Skills (UKCES) and (subsequently) the Department for Education (DfE) carried out a biennial Employer Skills Survey (ESS), which provides insights about the skill problems employers are facing to fill their vacancies and the actions they are taking to solve them. The survey contributes to public policy decisions when addressing the skills challenge and prompting people to adopt relevant skills for the workplace (Vivian 2016).

Another example in the UK is the Local Economy Forecasting Model (LEFM), developed by Cambridge Econometrics (CE) in collaboration with the Institute for Employment Research at the University of Warwick (Cambridge Econometrics 2013). Based on the 2011-based Sub-National Population Projections (SNPP) developed by the Office for National Statistics (ONS), and assuming that the historical relationship between growth in the local area and the region or the UK economy will hold in the future, this model allows researchers to project/anticipate different economic scenarios (skill forecast), as well as to evaluate possible skill mismatches at occupation or qualification levels, among other outcomes (Cambridge Econometrics 2013).

[34] They can be enumerated as follows: percentage change of median real pay (1 yr); percentage change of median real pay (3 yrs); return to occupation; change in median vacancy duration (1 yr); vacancies/claimant count; percentage change of claimant count (1 yr); percentage change of employment level (1 yr); percentage change of median paid hours worked (3 yr); change in new hires (1 yr); skill-shortage vacancies/total vacancies; skill-shortage vacancies/hard-to-fill vacancies; and skill-shortage vacancies/employment.

Moreover, reports such as "The Future of Work: Jobs and Skills in 2030" interview experts (senior business leaders, trade union representatives, education and training providers, policymakers, academics, etc.) from different sectors and conduct a comprehensive literature review, workshops, among other researches, to analyse sector trends and examine future economic scenarios (possible skill mismatches) and their implications for the labour demand for skills in the UK (skill foresight). These kinds of prospective labour studies are valuable because they estimate future employer requirements and address the education and VET system according to possible future needs using different and robust sources of information (UKCES 2014).

Other valuable efforts include the O*NET system launched in 1998, which is updated by the US Department of Labour, and the European Skills, Competences, Qualifications and Occupations (ESCO) in Europe, which is updated under the jurisdiction of the European Union. Based on the US Standard Occupational Classification (SOC) system, the O*NET system periodically consults a variety of resources—such as a national sample of establishments and their workers, occupational experts and analysts, among others—to collect information on hundreds of standardised and occupation-specific descriptors, e.g. knowledge, skills, tasks, work activities, and other descriptors (National Research Council 2010). Consequently, the O*NET provides an updated and detailed description of requirements for each occupation (skills, tasks, knowledge, etc.). With this valuable information, government officials can understand ongoing changes in the nature of work and their implications on the US workforce. Moreover, the O*NET identifies specific groups of occupations, such as "Bright Outlook occupations" or, in other words, occupations that are expected to grow swiftly in the coming years (potential skill mismatches) or will have considerable numbers of job vacancies. Consequently, this system helps the government to develop and train the workforce depending on their skill needs.

In addition, the Cedefop has made important advances towards quantifying skill needs in Europe. For example, the Occupational Skills Profiles (OSP) approach aims to integrate and complement several European sources of skill requirement information in order to provide updated occupational profiles for the region (Cedefop 2012b). Importantly, as mentioned above, the European Commission has built the ESCO, a multilingual classification system, which attempts to cover all European skills, competencies, qualifications, and occupa-

tions. It is important to note that occupations in the ESCO follow the structure of the International Standard Classification of Occupations (ISCO-08) at the four-digit level, and that the ESCO provides lower levels of disaggregation of skills for each occupation, such as an exhaustive list of 13,485 relevant skills (skills pillar) (European Commission 2017). This system was created to be compatible with other European platforms and supports an automated matching of jobseeker skills and vacancies. Consequently, in principle, the ESCO can be used to identify mismatches between CVs and vacancies in Europe.

3.5. Lack of accurate information to develop well-orientated public policies

In contrast with the above-mentioned classification systems in the US and Europe, Colombia does not have these kinds of advanced tools to base its education and training policies on them (González-Velosa and Rosas-Shady 2016). There exist some approaches to analyse the labour market in terms of skills, but there is not an integrated information system for skill mismatch analysis (Saavedra and Medina 2012). Institutions that have tried to measure, directly or indirectly, human capital characteristics have used different statistical approaches and skill concepts.

Since 2006, the Colombian statistics office (DANE) carries out a monthly cross-sectional household survey, the GEIH, to measure the characteristics of the Colombian workforce. The GEIH is nationally representative and constitutes the main source for official labour market information in Colombia. For instance, based on the GEIH, each month the national government publishes the unemployment rate and other relevant labour market indicators for Colombia. In this survey, people are asked about their current level of education and occupation, among other characteristics. As pointed out in Chapter 2, the level of education and the occupation of the labour force are two of the most common indicators to measure skill levels in a country.

However, for the Colombian case, this occupational analysis is limited for two reasons. First, the country's occupational classification system has not been updated since 1970. The DANE uses in its household surveys the Standard Occupational Classification (SOC), which was established in 1970 by the

Ministry of Labour and Social Protection and the SENA (Servicio Nacional de Aprendizaje), the vocational education and training institution in Colombia (Cabrera et al. 1997). The use of such outdated classifications might distort any subsequent statistical analysis due to labour market changes and new occupations that emerge or disappear over time. Occupations related to Big Data technologies (machine learning engineers, data scientists, and Big Data engineers) are representative examples, as these kinds of "Big Data" occupations did not exist 50 years ago, yet nowadays these are one of the top emerging jobs on LinkedIn (LinkedIn Economic Graph 2018).

Second, for analysis, the occupation variable is aggregated to two digits, which means that, for statistical purposes, the DANE aggregates the data into an "occupational area," which groups different occupations together depending on their qualification level (defined by the complexity of their functions, their level of autonomy and responsibility, as well as their level of education, training, and experience) (Sánchez Molina 2013). However, as mentioned in Chapter 2, the human capital concept has evolved and encompasses different elements—for example, socio-emotional, higher-order cognitive, basic cognitive, technical skills, among others—that are relevant for the labour market and cannot be measured using an outdated and aggregated classification system. Consequently, occupational data from the GEIH is useful as it provides insights about the general labour market structure, but it does not convey detailed information about skills and important human capital characteristics in order to develop national or local public policies on human resources.

In 2012, the World Bank carried out the STEP Skills Measurement Program to measure skills in low and middle-income countries, which included Colombia (Pierre et al. 2014). This program consisted of a longitudinal household-based survey and an employer-based survey. Nevertheless, for Colombia, only the household survey is available in which people were asked about (self-reported) personality, behaviour, and time and risk preferences, among other personal characteristics, and which also measured reading proficiency and related competencies according to the Programme for the International Assessment of Adult Competencies (PIAAC) scores to allow international comparison. Questions regarding skills make the STEP survey instruments a valuable source of human capital information in Colombia. The survey sought to be representative for non-institutionalised people from 15 to 64 years of

age, living in private dwellings in the thirteen major urban areas of the country. However, the general sample is composed of only 9,960 people, and after a short questionnaire, a member of the household was randomly selected to answer a more detailed individual questionnaire, which contained questions regarding skills. The total number of people who answered the skills modules is about 2,617 (Pierre et al. 2014).

Consequently, one of the main limitations of the STEP surveys is its sample size; indeed, it only represents 0.02% of the target population. Thus, the data sample cannot be disaggregated into different levels (i.e. different occupations) to make national or regional inferences due to the lack of observations. Additionally, the survey has not been updated: the first wave of information-gathering was conducted in 2012, and the second wave in June 2014; however, Colombia was not part of the second wave.[35] Therefore, as noted by the OECD (2017b), the STEP approach can be used as an instrument to understand some of the general structure in the skills performance of people aged between 15 to 64 years old in each country, and allows international comparison, especially with OECD countries. However, as the labour market is dynamic and skills performance changes over time, the survey needs to be updated, at least for the Colombian case.

Additionally, both the GEIH (DANE) and STEP (World Bank) surveys are based on what people (labour supply) report. Consequently, they do not directly consider one essential part of the labour market: employer requirements. An analysis of labour demand based on what people report in household surveys is limited because it only takes into account the skills or characteristics that people possess for the labour market, but employer requirements (what is needed to fill their vacancies) remain unknown, which is an important aspect of the labour demand to understand in order to reduce possible mismatches (Autor 2001; Mavromaras et al. 2013).

The DANE carries out sectorial surveys (e.g. industrial, services, and sales-retail activities) to measure basic information, such as national account statistics, the composition of production and consumption lines, the amount of labour employed in each sector, among other indicators. Subsequently, these

---

35 The following countries were included in the second wave: Armenia, Georgia, Macedonia, and Kenya.

surveys are not designed to obtain detailed information about human capital such as occupational structure, nor about the skills required for each position. For example, regarding human capital characteristics, with these sectorial surveys it is only possible to distinguish the number of people employed by different functional areas (e.g. production, marketing and sales, investigation and development, among others). Additionally, in 2012, the DANE carried out another cross-sectorial survey named the Human Capital Formation Survey, where companies in the above mentioned three sectors were asked about job training and productivity. Although the EFCH provided valuable insights about job training, selection and hiring practices, as well as productivity, the data are still aggregated by functional areas and do not capture employer requirements.

For its part, the SENA—the institution in charge of delivering vocational education and training in Colombia—also conducts small, voluntary employer surveys (semi-structured survey questionnaires) in order to identify the occupational requirements of the private sector. However, González-Velosa and Rosas-Shady (2016) argue that these surveys do not have enough financial resources to guarantee the effectiveness of their results. Indeed, the same authors highlight that employer survey results are significantly affected by a lack of standard procedures, clarity in their objectives, and incentives for companies to participate.

In 2015, the SENA surveyed both employees and employers in order to create employability, performance, and relevance indices of its vocational programs. The SENA sought to evaluate the skills performance of its graduates, such as communication, adaptation to changes, responsibility, teamwork, among others. Around 4,502 people were interviewed, who graduated from the institution in the second semester of 2013 and in the first semester of 2014, in addition to employers who hired those graduates (SENA 2015). The survey attempted to evaluate the content of vocational programs by measuring skills performance in people's jobs. However, even for that purpose, the results of these surveys are limited. Indeed, they are representative of only 13% of the total number of vocational programmes (SENA 2015), and employers were not asked about their skill requirements to fill vacancies. Moreover, information from the SENA (microdata) is not available to the public.

Thus, in Colombia, the main sources of information used in the analysis of labour demand have come from sectorial (entrepreneur) surveys or household

surveys. These data have certain strengths, such as national standardisation and global representativeness, but the collection of labour demand information through surveys is limited, since it can be quite costly, both in terms of resources and time. Above all, these sources might not provide enough detailed information about what skills (or occupations) are in demand among different industries or regions (Handel 2012; OECD 2016b).

In 2013, a Colombian law[36] established that all job portals and companies must report their vacancies to the Unidad Administrativa Especial del Servicio Público de Empleo (UAESPE, for its acronym in Spanish). Thus, potentially, the UAESPE can provide a vacancy data analysis for Colombia. However, the UAESPE approach has different limitations that affect the robustness of vacancy analysis. First, job portals and companies do not report all the information that describes a vacancy to the UAESPE. There is a predefined format that companies and job portals complete with certain information that partially describes the vacancy. Second, the UAESPE does not know whether companies report the total number of vacancies available. For instance, employers might underreport the number of vacancies because it might be time-consuming to fill and send all the information to the UAESPE. Moreover, the UAESPE does not have a methodology to systematically verify that employers have reported the total number of job vacancies advertised. Third, the inclusion or exclusion of some employers or job portals over time might affect the vacancy time series. An increase in the number of vacancies might be due to the inclusion of a new job portal (with not necessarily different and new vacancies). Fourth, as will be discussed in more detail in Chapters 4 and 5, the problem of duplication increases by adding more websites. The UAESPE collects information from different job portals and employers. However, a job vacancy can be published on various websites. Given that employers are not required to report full vacancy details, it is more difficult for the UAESPE to determine whether a vacancy is duplicated. Finally, the database and the UAESPE methodology to compile, clean, and classify vacancies are not available; hence, the vacancy analysis conducted by this institution lacks robustness.

These problems have made employer requirements or vacancy information scarce (Allen, Levels and Velden 2013). As Álvarez and Hofstetter (2014)

[36] Decreto 2852 de 2013.

mention, vacancy data to study the labour market are scarce in developing countries like Colombia. As a result, the human resource needs of the country have remained unknown until the present study. As a consequence, Colombia lacks a human capital formation system with accurate tools (among other instructional agreements) to address public policy, education and job training programs. So far, these aspects have remained unaddressed and have not been aligned with the employers' needs; instead, a low-quality education has proliferated. For instance, in 2013, only 4% of 1,576 technological training programs, and 3% of 740 professional technical training programs offered by private institutions were accredited (considering content and infrastructure, among other characteristics) in terms of quality by the Ministry of Education (González-Velosa and Rosas-Shady 2016). Likewise, the Regional Centres of Higher Education (CERES for its acronym in Spanish) have been reported to teach their students with outdated technologies and at an insufficient educational quality level (OECD 2016b). Given the low standards of training and education quality, even the Technical and Vocational Education and Training (TVET) system has not grown enough in the last few years due to lost prestige (OECD 2015b).

Given these facts, it has become necessary to seek new and novel ways to assess the labour supply needed by companies. One promising approach to address this issue is the provision and analysis of detailed labour demand information using Big Data techniques. As will be discussed in the following chapter, building a web-based model of skill mismatches (skill shortages) for Colombia (and potentially for its regional counterparts) might have a large impact, considering its potential use as a public policy tool related to a better management of human resources (i.e. the reduction of informality and unemployment rates), and also to assist in the allocation of skill development and educational budgets.

3.6. Conclusion

Despite the socio-economic improvements of the last decades, the Colombian labour market faces important challenges. The proportion of people participating in the labour market has considerably increased since 2008. Therefore,

the labour market needs to 1) engage new job seekers into the formal economy, 2) retain workers in the formal economy, and 3) move informal workers into the formal sector.

While other countries have created systems with statistical tools in order to measure skill mismatches and, thus, orientate public policies that seek to decrease this phenomenon, different barriers might prevent the pursuit of that goal in Colombia. According to the evidence presented in this chapter, skill mismatches are one of the most important barriers to reduce unemployment and to increase employment in the formal sector; consequently, skill mismatches might explain the high incidence of informality and unemployment in Colombia. A revision of the most important sources of information regarding human capital in the country shows that 1) available information sources are aggregated at levels that do not enable a detailed knowledge of existing occupations or skills; 2) there are difficulties in updating surveys or classifications (e.g. SOC 1970); 3) there are representation problems in the data gathering process (e.g. limited sample sizes), and 4) no information sources collect employer vacancy requirements. Thus, the available data indicate that there is a skill mismatch problem, which means that it is not possible to know in enough detail what skills are needed in the Colombian labour market.

The above analysis, in combination with institutional efforts, evidences the interest of Colombia in measuring and tackling skill mismatches. However, the absence of an accurate tool to measure the multiple dimensions of human capital, together with an institutional disarticulation, is one of the most critical factors that complicate the design of public policies, policies that need to be well-oriented in order to reduce the skill mismatch phenomenon in the country. Thus, a web-based model of skill shortages might provide valuable information for policymakers about employer requirements and might connect various efforts made by different institutions regarding skill mismatch analyses.

4. The Information Problem: Big Data as a Solution for Labour Market Analysis

4.1. Introduction

"More and better data" is a common claim of researchers and policymakers as a prerequisite to design public policies such as tackling skill mismatch issues (Cedefop 2010; OECD 2017b; Williams 2004). Information collection about labour demand through surveys involves statisticians, interviewers, and a sample of companies or individuals available and willing to respond. The cost of this kind of projects is relatively high, in terms of resources and time, and can discourage countries (especially with low budgets) from collecting and analysing vacancy data. Additionally, even if a survey is carried out, the information obtained might not be detailed enough to analyse which skills or occupations are in demand among different industries or regions (Handel 2012; OECD 2016d).

Currently, with the proliferation of the internet and higher-capacity electronic devices, large amounts of information about the behaviour of different agents are being stored daily. The storage of all this information has unlocked new territories for research in different areas of knowledge. For instance, Edelman (2012) and Askitas and Zimmermann (2015) detail several research examples using Big Data that have provided different applications for research in micro- and macroeconomics, labour and demographic economics, public economics, health, education, and welfare, among others.

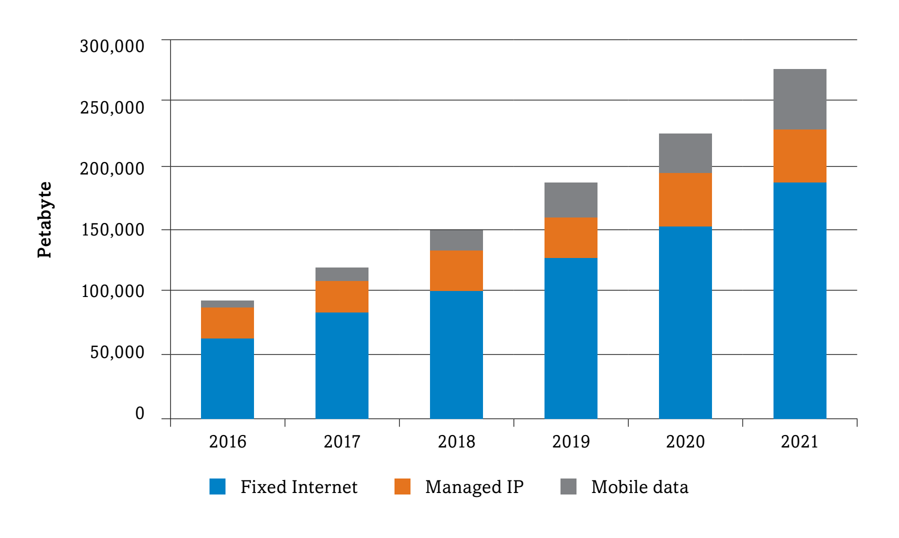
Big Data may be a way to overcome the limitations of existing skills analysis. More specifically, online job portals are a promising source of valuable information about labour demand. Thus, the second section of this chapter defines Big Data, followed by an analysis of how Big Data might fill information gaps in labour demand and supply to address labour market policies and research. The fourth section discusses the potential uses of information from job portals to tackle skill mismatches (skill shortages). Big Data in specific job portals has certain limitations and, for this reason, the fifth section discusses these limitations and indicates some caveats when using this kind of data

for analysing the labour market. Finally, the chapter describes how Big Data sources might facilitate the analysis of the labour market in a context such as the Colombian economy.

4.2. A definition of Big Data

Higher internet speed, a broader use of smartphones, tablets, cameras, computers, etc., and technologies with increasing capacities to store information have favoured the creation and storage of computerised or digital information on a large scale. Cisco (an important multinational technology conglomerate) estimates that 96 exabytes (1 EB = 10^{18} bytes) were the average monthly data traffic in 2016 and it is expected to increase three times by 2021 (278 EB per month) (see Figure 4.1). This era of massive information has unlocked opportunities for private and public institutions to compile, link, and analyse relatively large flows of data produced by different sources in order to better orientate important decisions and strategies. This set of massive information, including the techniques used to process and analyse the available information, is commonly labelled as "Big Data."

Figure 4.1. **IP traffic by source, 2016-2021**



Source: Cisco (2017, p. 6).

However, there is still an extensive debate about what can or cannot be considered as Big Data. Perhaps one of the most common conventions defines this term according to three properties: volume, variety, and velocity (Laney 2001). Each of these properties will be discussed in turn. The first one refers to the most obvious property to be considered as Big Data: the size (or volume) of data. In a simple way, data with a large volume of information might be a candidate to be called Big Data. However, individuals might consider different volumes of data differently because there are different computer capacities available in the market (with more or less data storage capacity, processing, etc.), which allow people to handle a certain number of bits per second. Consequently, it is necessary to determine a standard threshold to classify data according to their size. One way to do this is by classifying data whose size represents a challenge to be processed and analysed within the average range of computer technologies available as "big."

Nevertheless, it is important to keep in mind that the threshold to determine whether data have a high volume of information might change over time. Average computer capabilities increase over time; as technology improves, so does its capacity to process a high volume of information. Hence, what was considered as Big Data when this research started might get altered by the time this book is finished. Despite the changing nature of data, this criterion is still useful because volume allows researchers to distinguish between data sources in a technological environment that is constantly changing.

Variety refers to data structure. Unlike the information that comes from surveys, information from Big Data might not possess a well-defined structure to organise different variables in specific spaces (columns) within a database. Instead, the information might come from a range of unstructured or semi-structured sources and in different formats, such as social media, sensors, websites, mobiles, videos, etc. (Aguilar 2016). This characteristic makes data processing a challenge. Algorithms need to be developed to identify patterns (such as tags, keywords, among others) to obtain meaningful information. Thus, it is essential to consider that the concept of Big Data is not just related to volume; this concept also includes complex data qualities that make it necessary to have access to a higher capacity to store, process, and analyse the gathered information.

Finally, velocity refers to the speed at which data are generated. Nowadays, information is generated in seconds; people can share an opinion to thousands through platforms such as Twitter or Facebook and generate different reactions in an instant. Likewise, companies can post their current vacancies in real-time on various websites to quickly attract potential workers. This speed presents a challenge and an advantage for data processing and data analysis.[37]

For the purpose of this book, "Big Data" is considered as a relatively high volume of information, which is produced in a relatively fast way by different unstructured or semi-structured sources, and might be available in diverse formats, where the above mentioned three characteristics of volume, variety, and velocity make information processing and analysis processes a challenge *per se* with average technologies available in this given moment (in 2017).[38]

Despite many challenges, Big Data is expanding or opening a new frontier of knowledge (Askitas and Zimmermann 2015; Edelman 2012). Indeed, Big Data might fill information gaps that exist in different fields and regions where information to carry out well-oriented public policies was frequently scarce in the past (Azzone 2018). In the particular case of the labour market in Colombia, this information might give insights about the characteristics of the labour supply; and, more importantly, due to the general scarcity of labour demand data (especially in countries like Colombia), Big Data offers the possibility of having, for the first time, a detailed picture of employer requirements in real-time. The following section discusses in more detail how Big Data has provided new valuable information to analyse the labour market in different areas.

---

37 There are cases where information is not quickly generated (e.g. on a daily basis); nevertheless, they (e.g. medical records) might be considered as Big Data given the size of the database, which overpasses the current average computer capabilities.

38 The debate about what constitutes Big Data is still open. Özköse, Ari and Gencer (2015) or BBVA (2018) add other characteristics such as "veracity" and "value" to the Big Data concept. The former refers to the trustworthiness or credibility of data, although this is an implicit characteristic that any data should have; while the latter term establishes that information needs to provide some profit (usually measured in terms of money) to a certain institution. Nonetheless, not every institution or person seeks monetary profit from information. For instance, non-profit institutions might benefit from Big Data information in order to provide goods and/or services for free or at prices that are not economically competitive. Moreover, the value of information depends on the observer: data that might not produce any value for a company might hold some value for a researcher or a different institution.

4.3. Big Data on the labour market

"Good" data are a requisite to develop well-oriented policies and academic research, where "good" refers to data that involve analysing the representativeness of the population and, thus, a minimum standard of quality during the collection process. Supply and demand information from surveys has limitations that Big Data might alleviate in order to form a better picture of the labour market. Thus, different efforts have started to be developed from both the supply and demand sides in some countries and areas, which involve the usage of Big Data.

4.3.1. Labour supply

4.3.1.1. Household surveys for the analysis of the labour supply

Traditionally, on the labour supply side, information has been collected from household surveys (e.g. employment rates by age, region, gender, etc.). Generally, these household surveys are characterised by a sampling frame (based on a census) that is representative of a specific population, a set of questions, and flows that customise the sections participants complete. Such surveys collect the main characteristics of the labour supply over a certain period. In most cases, the surveys are carried out by the Office for National Statistics (ONS) of each country, which follow certain quality standards provided by an international institution such as the ILO. These standard procedures make household surveys one of the main sources of information to calculate indicators of the labour market, such as participation and unemployment rates, wages, etc.

Despite its indisputable advantages, household survey information has some limitations that might be overcome with Big Data. First, information collection through surveys requires time for design, validation, collection, and consolidation, among other processes, that might delay the publication of the resulting database for analysis. When data are available, the researcher needs time to process the information, to analyse possible alternatives, or to address specific issues. However, an important disadvantage of such methods is that some time will elapse from the moment when the survey is designed until the

completion of the final database, and during this time the analysis of data might become outdated and invalidate the research findings due to changes in the socio-economic environment. Indeed, Reimsbach-Kounatze (2015) highlights that many OECD countries only have access to labour supply information several weeks (at best) after the data were collected.

Second, another limitation of household surveys is their fixed structure as a pre-designed questionnaire, which collects information on a variety of topics from people for various monitoring, planning, and policy purposes. Surveys also have budget and time constraints. For instance, the UK Labour Force Survey (LFS) aims to measure "economic activity and inactivity, all aspects of people's work, job-search for the unemployed, education and training, income from work and benefits" (ONS 2015). Clearly, variables that are beyond this scope are not measured. Moreover, adding one single question increases the survey's cost and might also affect its structure, flux, response rate, and results. This makes it difficult for survey designers to include other relevant labour supply questions. Thus, household surveys are a rigid tool that attempt to measure social issues, whose dimensions might change over time.

Third, due to sample constraints, household surveys have a statistical limitation. The more the data are disaggregated (e.g. region, sector, age, education, etc.), the more imprecise are the estimates. For instance, the GEIH survey has available labour market results, such as employment or unemployment shares, disaggregated by city and SIC (Standard Industrial Classification, revision 3). This information is useful to analyse unemployment rates by region, major occupational groups, etc; nevertheless, the level of detailed information (granularity) obtained from household surveys might not be sufficient to cover topics, which might be particularly useful for institutions and individuals (e.g. sector employment composition, the skills possessed by individuals, and occupations).

Additionally, household surveys (and, in general, other kinds of surveys) are not exempt from limitations, such as measurement errors (the difference between a measured quantity and its true value), issues when collecting the information (e.g. interviewees might provide imprecise or false information or interviewers can make mistakes when recording the data, etc.). Thus, as stated above, household surveys have important limitations: 1) a time lag between designing the survey, data collecting and processing, and analysing the results; 2) a fixed structure that makes it difficult to include or modify questions and

update categories; 3) a design to be representative for a certain population at a disaggregated level; and 4) other potential limitations such as measurement errors and issues in data collection.

Consequently, several variables of interest for policymakers and researchers are not provided by household surveys. Such is the case of job networking, among other job seeker behaviours. Therefore, although household surveys are one of the main sources of labour supply information, there exists relevant uncovered information, which might be provided by using Big Data information.

4.3.1.2. Big Data and labour supply

So far, the contribution of Big Data information to knowledge about labour supply has come from two sources. The first source uses search engines such as Google Insights, and the second source uses social media and networking sites to monitor (over a relatively short period) the behaviour of job seekers. Regarding the former, search engines track millions of searches in real time concerning different topics, such as weather, news, products, and, importantly for this document, job searches. Consequently, these word searches can be used to identify trends in people's behaviour. For instance, Askitas and Zimmermann (2009, p. 6) found that the usage of certain keywords—for example, "unemployment office or agency," "unemployment rate," among others—by German people on Google has a strong correlation with, and therefore is a useful predictor of, the unemployment rate in Germany. The underpinning idea is that people will use certain words related to job searches in Google when they are (or are likely to be) fired, or when it is difficult to find a job. Thus, access to people's searches on these kinds of search engines can provide information before the results of official surveys are available.

Regarding the latter, social media and networking sites might be a source of labour supply information. Specialised social media platforms and websites, such as LinkedIn and BranchOut, have arisen in the last decade. For instance, LinkedIn is one of the most well-known professional networks as it is present in more than 200 countries and has more than 552 million users (with around 250 million users active every month), who make their curriculum vitae public in order to be contacted or contact potential employers (LinkedIn, n.d.). The information available through these social media platforms might provide

insights about the skills and other characteristics of the labour supply (see, for instance, State et al. 2014).

Interestingly, information from social media platforms has helped researchers to build or further refine their employment indicators. Such is the case for Antenucci et al. (2014), who created indexes of job loss, job searches, and job postings in real-time by tracking keywords such as "lost job," "laid off," and "unemployment", among others. Thus, regarding labour supply, social media and networking sites as well as search engines have created the opportunity for researchers to deepen their understanding in certain topics.

4.3.2. Labour demand

Perhaps the use of Big Data for labour demand analyses has raised higher expectations among researchers, policymakers, etc., than its use for labour supply. These expectations might be motivated by the fact that, traditionally, labour demand information has been scarcer than labour supply information (Kureková, Beblavy, and Thum, 2014). As explained in more detail in this section, labour demand information shares many of the same limitations as labour supply information, such as sample constrains and granularity. However, unlike labour supply information, labour demand surveys and the analysis of employer requirements tend to be less frequent (especially in countries like Colombia; see Chapter 3). Paradoxically, as Hamermesh (1996) emphasises, one reason that explains why studies about labour demand have been relatively ignored or scarce is that the "creation of large sets of microeconomic data based on household surveys has spurred and been spurred by development of new theoretical and econometric techniques for studying labour supply" (p. 6).

Consequently, the main sources of information used for the analysis of labour demand have come from sectoral surveys (such as industry surveys) or even from household surveys. Even though these data sources have strengths, such as national standardisation and representativeness, the collection of labour demand information through surveys is likely to be costly, both in terms of resources and time, and these surveys might not provide enough information to workers, governments, and other institutions about human resources needs.

4.3.2.1. Sectoral surveys

In the UK's "Vacancy Survey," carried out by the Office for National Statistics (ONS), around 6,000 trading businesses[39] are interviewed monthly to provide "an accurate and comprehensive measure of the total number of vacancies across the economy and fills a gap in the information available regarding the demand for labour" (ONS, n.d.). The survey's main results are published in the Labour Market Statistical Bulletin within six weeks of the reference date of the survey and reveal the monthly number of vacancies in the UK. Additionally, there is a time series available regarding the total number of vacancies (seasonally adjusted) by industry, which are aggregated (following SIC 2007 sections, 22 groups) by the size of businesses[40] (ONS 2016a; ONS 2016b), and a time series comparison between the total number of vacancies and the total number of unemployed people (Beveridge curve) (ONS 2016c). Moreover, the UK Employer Skills Survey (carried out by the DfE) provides detailed information about job requirements; specifically, skills and occupations demanded by employers (at the four-digit SOC level if possible), and industries (22 major groups at a one-digit level according to SIC 2007). This survey is a biennial study, and its main results are published over the months following each survey (Vivian 2016).[41]

Likewise, less developed regions like Colombia have made different efforts to collect and analyse labour demand. As mentioned in Chapter 3, Colombia has conducted sectoral surveys, such as the annual industrial survey and services survey. Despite these considerable efforts, these kinds of sectoral or cross-sectorial surveys (e.g. EFCH[42]) present severe limitations for the analysis of labour demand and, consequently, skill mismatches. First, as the name suggests, sectoral surveys are applied for a specific sector. The EFCH survey

---

[39] It excludes agriculture, forestry, and fishing.

[40] 1-9 employed; 10-49 employed; 50-249 employed; 250-2,499 employed; and 2,500+ employed.

[41] At the time this document was written, the last available report was published in 2015, and the results for the 2017 survey was set to be available in summer 2018 (ONS n.d.).

[42] To carry out this survey took an investment of $397,349 (from the Ministry of Labour and the Interamerican Bank of Development, IDB), plus the DANE had been providing survey preparation in terms of sample designs, logistics, and advice since 2010 (CONPES 2010; DANE 2018c).

in Colombia is only applied to companies related to industrial services and sales-retail activities. Therefore, some sectors might be excluded, and their labour demand composition and dynamics will remain unknown. Second, not all types of companies are included in the sampling frame. The industrial EFCH survey interviews establishments with 10 or more employees and whose annual production is above £125,000. Moreover, the EFCH survey's results are available according to "functional areas" such as "Production," "Management," and "R&D." Thus, these sources might not be enough to provide detailed information about which skills (or occupations) are in demand among different industries or regions (Handel 2012; OECD 2016d).

Likewise, the Colombian annual industry survey, which is one of the main sources for labour demand information, interviews establishments with the same criteria than the EFCH. Indeed, the EFCH is a subsample of the annual industry survey sample. Consequently, many companies (generally small or medium-sized companies) in a sector might not be included in the sample, so even within the relevant subsample, part of the labour demand is ignored.

Perhaps more advanced regions such as the UK are less exposed to this aggregation problem. With a greater budget, these regions can design surveys with a higher disaggregation level, as is the case of the UK Employer Skills Survey mentioned above. Nevertheless, even with a larger budget, the results from industry surveys might be produced with a relatively low frequency. For instance, the main findings from the UK Employer Skills Survey are released every two years. Policymakers, educational institutions, and researchers, among others, need to wait at least two years to access the information collected by the survey about labour demand requirements. There is a long period between the time the survey is carried out and data are processed, cleaned and released; labour conditions might change during the two years it takes to prepare data reports, and consequently, some results might be outdated. Regarding this problem, less advanced countries like Colombia are in a worse situation. For instance, in 2018, at the time this section was written, the last EFCH survey had been conducted in 2012 (a period of 6 years).

In some industry surveys, companies or a group of experts are asked about the number of vacancies that opened in the last period (e.g. within the last year), the number of vacancies that each company is expected to have in the next period (e.g. within the next year), the expected volume, and some

general characteristics (e.g. experience) of people that they will need in a certain period of time (e.g. the following three months, six months, a year, etc.). By providing information about current and future labour demand dynamics, such questions address the problem of the low frequency of data results.

Based on this labour demand information, two different approaches have been developed to anticipate future labour market needs: skill forecast and skill foresight. The first term refers to forecast exercises that "use available information or gather new information with the specific aim of anticipating future skills needs, mismatches and/or shortages. Forecast results are meant to provide general indications about future trends in skill supply and/or demand in the labour market" (OECD 2016d, p. 39). The latter term, skill foresight, aims to "provide a framework for stakeholders to jointly think about future scenarios and actively shape policies to reach these scenarios" (OECD 2016d, p. 39). Both these exercises are valuable because they estimate future employer requirements and address the education and VET system according to possible future needs.

Nevertheless, once again, efforts like skill forecast and skill foresight are relatively expensive in terms of money and time, and their results are too specific to be of use to the broader labour market. For instance, in Colombia, prospective labour market studies focus on specific sectors, such as coffee production and building construction. Moreover, projections from skill foresights or skill forecasts might be biased or mistaken. For example, companies might experience sudden expansion (or contraction) periods, which can unexpectedly increase (or decrease) the creation (or destruction) of future vacancies. Thus, labour demand estimates might under- or overestimate the number of vacancies and their characteristics. Likewise, experts might not accurately predict the course of a sector over the long term. Additionally, parameters to make economic projections might be outdated. Consequently, projections based on these data would ignore economic changes that have occurred between 2005 (date of the most recent census available at the time this section was written) and the date when a new census is conducted, and economic projections are re-estimated.

Therefore, sectoral surveys and exercises derived from them have several limitations: 1) They require large logistical operations and a substantial amount of money to conduct a labour demand survey. Consequently, 2) considerable time is needed to design, collect, process, and release the information gathered. 3) Given budget constraints and survey designs, some companies or sectors

might be excluded from labour demand analysis. For the same reasons, 4) it is, frequently, unlikely to be able to disaggregate survey results at numerous levels: occupational, skills, industry, region, etc. Given these limitations, labour demand information is scarce and less frequent (e.g. monthly) than household surveys. Finally, 5) skill forecast or skill foresight methods might not properly foresee economic changes and their implications for skills (labour demand). Therefore, it is relatively common to find labour demand studies in the economic literature whose main sources of information are household surveys.

4.3.2.2. Household surveys for labour demand analysis

Traditionally, household surveys have functioned as inputs to analyse labour demand issues. These sources provide information about the intersection between labour supply and labour demand (filled labour demand) over a certain period. Household surveys provide data about labour demand in the following way: employed people can occupy one or more job vacancies, consequently, the total number of employed persons weighted by the number of jobs held by each one of them is equal to the total number of vacancies filled (satisfied demand; see Chapter 2).

This information about satisfied labour demand has been used in different studies as an approach to analyse labour demand dynamics. Moreover, the availability of a relatively long series of household data has allowed analysing relevant trends and changes of the (filled) labour demand (Acemoglu and Autor 2011; Autor and Dorn 2012; Autor, Katz, and Kearney 2006; Salvatori 2018).

However, to analyse the labour demand based on what people report on household surveys is limited. First, as explained above, survey constraints (e.g. money and time) might not allow disaggregating the results at a skill or occupational level (e.g. 4-digit level ISCO). Second, household surveys only take into account the current/past skills or characteristics of the workforce; what is unknown are employer requirements to fill their vacancies, which is an important aspect of labour demand to reduce possible mismatches (Autor 2001; Mavromaras et al. 2013);[43] thus, the acquisition of information is based

[43] For instance (as mentioned in Chapter 3), the World Bank has conducted a Skills Measurement Program to assess skills in low- and middle-income countries (Pierre et al. 2014).

on what people (labour supply) report, and does not consider an essential part of the labour market: employer requirements.

This issue is an important limitation when considering employment share as a proxy of the labour demand. Total employment is at the intersection between labour supply and demand. Nevertheless, the level of employment might significantly differ from the true level of demand because of unfilled labour demand (vacancies). For instance, employers might demand high-skilled jobs, but there is no labour supply to fill them; consequently, by only using total employment, the fact that there is an important demand for high-skilled workers would be ignored.

Therefore, household surveys are a valuable input to analyse filled labour demand and its long-term changes. Nevertheless, this information is limited in the following aspects: 1) there are constraints (e.g. time and money) that affect the level of aggregation and the frequency of data collection; 2) these surveys do not capture information about employer requirements, which is essential to address issues such as skill shortages. Consequently, all the problems mentioned above for sectorial and household surveys restrict the capacity of researchers and policymakers to tackle skill mismatches.

4.3.2.3. Big Data and labour demand

As previously mentioned, the collection of labour demand information is relatively less systematic than labour supply information. Moreover, even when labour demand information is available, different limitations make skill mismatch analysis a challenge. However, it seems that the proliferation of high-volume information (such as the internet) and techniques to analyse it have brought the opportunity to evaluate possible skill mismatches (skill shortages) through the analysis of employer requirements.

Nowadays, internet is an important source of information. This source is widely used for different purposes, and it stores relevant information regarding the behaviour of agents such as employers. As Autor (2001) highlights, the internet provides an opportunity to collect more and possibly better labour market data. Indeed, online information contains a large number of detailed observations about labour demand, and it can be accessed mostly in real time and at a relatively inexpensive cost (Barnichon 2010; Edelman 2012).

Moreover, the use of the internet by employers to advertise and find suitable applicants, and by individuals to find a job, has dramatically increased. As mentioned by Maurer and Liu (2007) and Smith (2015), both employers and job seekers have increasingly used the internet to find a vacancy or to advertise. In fact, by 2007, more than 110 million vacancies and 20 million unique resumes were stored in online US sources (Maurer and Liu 2007, p. 1). More recently, Kässi and Lehdonvirta (2018) suggest that the volume of online new vacancies has grown roughly 20% worldwide from 2016 to 2018. Likewise, the number of job seekers looking for a job using online sources has increased. For instance, in the US, the share of people who used the internet to find a job increased from 26% in 2000 to 54% in 2015 (Smith 2015).

The use of online job portals as a source of information has grown among researchers and has also attracted the attention of policymakers because they seem to provide quick and relatively inexpensive access in order to analyse information about employer requirements. Job portals are websites where companies make public their current (or future) vacancies. Companies describe, to some extent, the job position and the attributes that a potential worker should have in order to be considered as a candidate. Additionally, job seekers can screen and select vacancies, and contact potential employers. In other words, job portals help to connect employers with job seekers and vice versa.

Information from job portals, however, is not produced for the purpose of economic analysis (indeed, in most cases, it is posted online by private businesses). Yet job advertisements can potentially function as an essential input to analyse the employers' needs. The systematic collection of information from job portals might help to diagnose the performance of an economy in real-time (e.g. at the level of available vacancies), and to understand employer requirements and how these requirements change over time. Consequently, along with the increasing usage of the internet and job portals, studies have used online job vacancy data to provide insights about the labour demand in different countries, such as the US, Slovakia, Czech Republic, and Colombia (Guataqui, Cárdenas, and Montaña 2014; Carnevale, Jayasundera, and Repnikov 2014; Marinescu and Wolthoff 2016; Štefánik 2012; Tijdens, Beblavy, and Thum-Thysen 2015).

In this sense, Kureková, Beblavy, and Thum (2014) have emphasised that job portals can be useful to generate a better understanding of company needs,

which might enrich labour market policies. In contrast with household surveys (filled demand), information from job portals (unfilled labour demand) might be useful to reveal what occupations or types of skills are currently in demand. Moreover, this kind of data might be of more relevance in contexts where employers experience difficulties to fill job vacancies, and information from job portals might be the only data available to analyse the labour demand for skills to address labour supply according to employer requirements. Consequently, in less advanced regions such as Latin America (e.g. Colombia), where the largest skill mismatches exist, there is a lack of labour demand information (see Chapter 3), and the usage of information from job portals to measure employer requirements might have a high impact on different labour demand outcomes.

4.4. Potential uses of information from job portals to tackle skill shortages

Targeted vacancy information gathered from online job portals might improve information and public policy deficiencies regarding skill shortage problems in the following ways: it allows 1) maintaining an estimation of vacancy levels, 2) identifying skills and other job requirements, 3) recognising new occupations or skills, and 4) updating occupation classifications.

4.4.1. Estimating vacancy levels

The number of job offers, together with other labour market indicators (such as unemployment levels), help to determine the business cycle and possible mismatches in an economy. High vacancy rates might mean that the economy is in a stage of economic expansion and/or there are mismatches between the supply and labour demand.[44]

In this sense, online job vacancy advertisements might provide real-time access to job offers in an economy, and public policymakers might react

[44] An example of the above is the Beveridge curve, which relates unemployment and vacancy to determine how well, or not, vacancies match with unemployed workers (Blanchard and Diamond 1989) (see Chapter 9).

or re-design public policies in a shorter period aligned to current economic changes. Given the advantages of collecting online information, different countries have started to create job vacancy databases based on information from the internet. For instance, in the US, there is the Help Wanted Online Data Series created by The Conference Board (Conference Board n.d.), and in Australia, there is the Internet Vacancy Index developed by the Australian Department of Education, Employment, and Workplace Relations (DEEWR) (Australian Government 2018). Both provide measures of labour demand (advertised vacancies) at various levels, including at a national, state, regional, and occupational level (Reimsbach-Kounatze 2015).

Moreover, online job vacancy information is not limited to counting the number of job offers in the economy. Indeed, one of the most important advantages of online job vacancy advertisements is that they provide detailed information about employer requirements. This aspect allows researchers, policymakers, among others, to delve into topics (that previously were relatively difficult or costly to obtain updated information on) and to identify demand for skills and other job requirements.

4.4.2. Identifying skills and other job requirements

Perhaps one of the most promising uses of online vacancy information is the identification of job requirements in a relatively short time to enable public policy design. As will be seen in more detail in Chapter 5, companies post their job vacancies on job portals along with detailed candidate requirements to fill each position (skills, education, experience, etc.). This detailed information creates an opportunity to monitor job requirements at a disaggregated level (e.g. 4-digits occupation level) and, for instance, advise VET institutions regarding what skills they need to train people in to increase their employability.

In this sense, one of the most important ongoing projects, at the time this book was written, is the "Big Data analysis from online vacancies" project carried out by the European Centre for the Development of Vocational Training (Cedefop for its acronym in Spanish). The Cedefop combines its efforts with Eurostat and DG Employment, Social Affairs, and Inclusion to collect data on skills demand using online job portals. With this information, the Cedefop attempts to monitor skills and other job requirements at an occupation level

and identify emerging skills and jobs in Europe to advise training providers to revise or design new curricula according to current labour demand requirements in Europe (Cedefop 2018).

Moreover, private companies such as Burning Glass Technologies provide and analyse labour demand information using job portals for countries like the US and the UK. For instance, this company has reported that 80% of middle-skill job advertisements demanded digital skills in 2016, which represents an increase of 4% compared with 2015 (Burning Glass Technologies 2017, p. 3).

4.4.3. Recognising new occupations or skills

As was mentioned in Chapter 3, labour market changes rapidly and new occupations or skills might emerge or disappear over time. The identification of these new patterns in labour demand is relevant because it allows curricula to be adapted by training providers and, as a consequence, it prepares people for technological change. Patterns in labour demand can be identified by recording labour demand information from job portals. For instance, Emsi, a labour market analytics company, has started to build a skill taxonomy, which has identified the growing demand for relatively new skills, such as "Cloud engineers/architects" and "Cloud computing" (Verougstraete 2018). Verougstraete (2018) mentions that this information might be useful to understand how to adapt the labour supply according to changes in labour demand—especially for the most innovative sectors such as IT and tech.

4.4.4. Updating occupation classifications

With a demand for identification of occupations and skills, and new emerging patterns for job requirements, job portal data might facilitate the updating of occupation classifications with real-time information. As mentioned in Chapter 2, occupation classifications are usually not updated as fast as labour market changes occur. A significant amount of time and financial resources are required to analyse information collected from companies and other stakeholders to update an occupation classification. However, with the relatively quick and inexpensive collection of online job advertisements it is now possible to identify job requirements (skills, educational level, tasks, etc.) of each

occupation; hence this information might become an essential contribution to update occupation classifications according to changes in labour demand.

For instance, as recognised by the ILO (2008, p. 2), "some countries may not have the capacity to develop national classifications in the short to medium term. In these circumstances it is advisable for countries initially to focus limited resources on the development of tools to support implementation of ISCO in the national context, for example a national index of occupational titles." In these circumstances, online job advertisements might provide relevant information to adapt ISCO classifications to a regional context.

Consequently, information from job portals can be used for a range of different topics. Authors such as Turrell et al. (2018) use job vacancy information to understand the effects of labour market mismatch on UK productivity. Moreover, Rothwell (2014) employs advertisement duration as a proxy of vacancy duration in order to determine skill shortages in the US. Additionally, Marinescu and Wolthoff (2016) and Deming and Kahn (2018) use online job advertisements to determine the portion of wage variance explained by employer skill requirements (e.g. cognitive, social, writing, etc.) in the US. However, one of the most promising uses of this information is the identification of skill mismatches. The study of labour demand for skills is a key input to overcome informational barriers between labour demand and supply (Kureková, Beblavy, and Thum, 2016). Yet, as the next section will address, despite the potential of vacancy information, it is essential to take into account its possible limitations, so as to avoid potential biases when analysing information from job portals.

4.5. Big Data limitations and caveats

It is important to note that despite the advantages of Big Data, such as the greater volume of information it allows researchers to analyse, there exist some limitations that might affect the analysis of labour demand via information from job portals. Consequently, any study that uses online job advertisements should consider the following issues: 1) data quality; 2) job postings do not necessarily represent real jobs; 3) data representativeness; 4) internet penetration rates, and 5) data privacy.

4.5.1. Data quality

Data quality is one of the most important factors that determines the reliability of any database for statistical purposes. According to the quality framework and guidelines provided by the OECD, data quality is a multi-faced concept within which the relative importance of each dimension depends on user needs. These dimensions are relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence (OECD 2011, pp. 7-10) (Table 4.1).

Table 4.1. **OECD quality framework and guidelines**

Criteria	Description
Relevance	"Degree to which the data serve to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concepts."
Accuracy	"Degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure."
Credibility	"Refers to the confidence that users place in those products based simply on their image of the data producer…This implies that the data are perceived to be produced professionally in accordance with appropriate statistical standards, and that policies and practices are transparent. For example, data are not manipulated, nor their release timed in response to political pressure."
Timeliness	"Reflects the length of time between their availability and the event or phenomenon they describe but considered in the context of the time period that permits the information to be of value and still acted upon."
Accessibility	"Reflects how readily the data can be located and accessed."
Interpretability	"Reflects the ease with which the user may understand and properly use and analyse the data. The adequacy of the definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any, largely determines the degree of interpretability."
Coherence	"Degree to which they [data] are logically connected and mutually consistent."

Source: OECD 2011.

With regards to these conditions, given the nature of Big Data on specific job portals as sources of information, this source has a clear advantage in terms of "timeliness" compared with other sources of information such as sectoral surveys. However, as mentioned in Subsection 4.3.4, job portals and, in general, Big Data sources (such as LinkedIn) were not initially created for policy

or academic purposes. This makes the data available through these websites seem relatively disorganised; for example, without standardisation, with duplication issues, and/or with a relatively high portion of missing values. Hence, data quality and the analysis of labour demand using Big Data sources might be affected or limited by these issues of organisation. Table 4.2 lists possible problems that might affect the quality of information provided by job portals.

Table 4.2. **Possible sources that affect the quality of information from job portals**

Potential data quality issues	Description
Employers do not follow a specific format when they advertise vacancies	For instance, where the online content indicates the presence of a "job title," there may also be information regarding the company's name, location, etc. This unstructured way of announcing vacancies can make statistical inference difficult. For instance, to generate a simple tabulate of a specific variable (e.g. wages), it is necessary to first identify where all (or most) of the information is located on the website and put only this information together to form the corresponding tabulate. Moreover, companies use their own "language" when providing information, such as job descriptions, titles, and the required skills; thus, employers might use different words to define a similar job position (see Chapter 5).
Companies are not required to provide a standard set of detailed information about the vacancy	The high occurrence of missing values might create bias in the analysis of a certain database. For instance, employers might reveal the wages offered for low-skilled jobs while they might not reveal the wages offered for high-skilled jobs. In consequence, when the mean of wages offered are estimated from any subsequent database, the results would underestimate the average of real wages due to missing information of a specific (high-skilled) occupation group (see Chapter 6).
Duplication issues	There are two possible types of duplication: within and between job portals. The first type ("within") refers to the situation where companies might advertise the same job position on the same job portal more than once. The second type ("between") occurs when employers advertise the same vacancy on more than one website. Consequently, when collecting information about labour demand using different job portals, the number of job vacancies might be overestimated, hence any statistical inference might be biased (see Chapter 6).
Mistakes in the information	Employers might make mistakes when typing in information, and, in some cases, the information provided might be contradictory. For example, when an employer writes in the job description that work experience is not required, but in the job title it states that some work experience is required (see Chapter 6).

Source: Author's elaboration.

All the problems cited above show that, when working with information from job portals, important issues need to be addressed to guarantee a certain level of data quality (some of these issues are also true of survey information). Clearly, the problems mentioned above can be reduced with the use of data mining techniques, such as data cleaning, classification, and imputation, among others, but they might not be completely eliminated. This result depends on the effectiveness of the algorithms used and the information provided by the employer (Chapter 10 discusses whether the vacancy database for Colombia fulfil the quality requirements established by the OECD).

Thus, the level of these data quality problems and the techniques implemented to tackle them will determine the extent to which information from job portals can be used to analyse labour demand. However, data quality is not the only concern when this type of information is used for analysis. There are other issues: job postings might not necessarily be real jobs, data representativeness, internet penetration, among other issues, might limit the usage of Big Data for the analysis of labour demand.

4.5.2. Job postings are not necessarily real jobs

Given the nature of job portals, any company or individual can post a vacancy.[45] However, job portals do not have the means to verify whether the advertisement corresponds to a real vacancy or might not be interested in doing so. As Sentz (2013) remarks, when using information from portals, there are difficulties in making a one-to-one comparison between job advertisements and a real job vacancy. For instance, companies might post more job advertisements than available positions in order to receive more applications, and then hire the candidates who best fit their requirements. Another alternative is that companies (such as recruitment agencies) might advertise vacancies to collect CVs and store them in their databases. With this technique companies have already collected the data of potential workers and have the ability to quickly start the screening process in the eventuality of a job opening.

45 Depending on the job portal, advertising a vacancy might be free or associated with a cost that generally depends on the time the advertisement is active on the website.

If job portals can post jobs that are not real, or companies can open vacancies without posting them, it is then difficult to precisely determine the number of job vacancies for an occupation, sector, etc., using job portals. These issues do not mean that information from job portals cannot be used as a source to analyse labour demand. With this information in mind to utilise the proper statistical techniques, it is possible to comprehend the structure and trends of labour demand (see Chapter 8); although it may be challenging to determine the exact number of real vacancies available in a period through information from job portals.

Moreover, as Sentz (2013) discusses, even with the above problems, job vacancy advertisements are useful to understand current skill demands, such as who is (or interested in) hiring and where the most employee rotation (turnover) is occurring. For instance, an employer might advertise ten job positions for accountants in a single job advertisement when he/she will eventually only hire five of them. Despite the possibility that job advertisements might overestimate the number of available vacancies, this information might reveal occupations and the demand for skills associated with those occupations. Therefore, information from job portals is a valuable resource to support the analysis of labour demand, even if not all advertisements correspond to a real job position.

4.5.3. Data representativeness

Even though information from job portals contains a considerable amount of data, this does not guarantee that this information is representative of the whole economy. On one side, some companies with a specific characteristic (sector, localisation, etc.) might not commonly use job portals to advertise vacancies. On the other side, even in the unlikely situation that every company uses job portals, some specific job positions might exist that are not advertised on websites. For instance, companies might recognise that people with low skills do not tend to use the internet to find a job, and the most effective way to recruit such candidates is through informal channels, such as one-to-one or personal references (e.g. friends). In consequence, depending on the available information on job portals, in some cases, it is not possible to make any statistical inference for a labour market segment or, in other cases, there might be some restrictions when the data are analysed.

Thus, when using information from job portals, it is relevant to understand which segments of the market are properly represented by these sources of information. This discussion of data representativeness is one of the main concerns regarding the use of this type of information for policy recommendation. The representativeness issue determines whether it is possible to analyse and make public policy recommendations for labour markets based on information from job portals. However (as will be discussed in more detail in Chapter 8), testing data representativeness is a complex task. To illustrate this point, it is important to consider how household surveys or sectoral surveys guarantee data representativeness. As mentioned in Subsection 4.3.3, household surveys are based on a population census. This census enables researchers to obtain information about the total number of individuals ("universe") and their main characteristics over a certain period. When the population and its characteristics are known, it is possible to draw a household sample. In this way, the information from household surveys guarantees that their sample results are as close as possible to the required population parameters (age, gender, etc.).

However, usually, in the case of vacancy analysis, the "universe" is unknown: for instance, the total number of vacancies available in a period by population groups (sector, occupation, localisation, etc.). Therefore, in this case, it is more difficult to know which population is represented by job portal sources. Paradoxically, the relative absence of vacancy information motivates researchers to use information from job portals; nevertheless, this absence of representativeness might limit or put in doubt the usefulness of job portal data.

Some authors such as Štefánik (2012) and Kureková, Beblavy, and Thum (2014) have addressed this issue. However, as pointed out by Kureková, Beblavy, and Thum (2014), most of the studies that have used job advertisements (printed or online) do not discuss or test data representativeness, and their findings are generalised for occupational or sectorial groups. The absence of discussion aimed at identifying data representativeness might affect the reliability of many studies. For instance, given the nature of the internet, occupations related to internet technologies (IT) tend to be overrepresented in online job advertisements. Consequently, a study that does not account for this source bias might conclude that IT skills are one of the most relevant skills required to find a job, while considering the total number of real vacancies (those advertised and not advertised on the internet), the actual share of IT occupations might be minimal.

Therefore, discussing and testing the representativeness of job portal data for academic and public policy purposes is a key issue when considering the use of these sources of information. The validity and the generalisation of results from the analysis of online job advertisements depend on the population being represented by job portal sources. For this reason, Chapter 8 discusses and tests data representativeness for the Colombian case.

4.5.4. Limited internet penetration rates

Related to the above point, the usefulness of information from job portals and, hence, their representativeness depends on internet penetration rates (the percentage of the total population that uses the internet). Although internet usage has increased (see Section 4.2), this growth might not cover some sectors, regions, etc. For instance, in Colombia, there is a remarkable disparity between rural and urban zones in terms of internet access.[46] Given this limited access, employers might tend towards the use of other job advertising channels such as asking friends or colleagues to recruit potential workers.

In regions where the growth of internet access has not occurred or has occurred at a slower pace, the inferences that can be drawn from job portal data might be more restricted than in areas where internet access is more widespread. Places where there is less internet access tend to be poorer, and information about labour demand tends to be scarcer due to the prohibited cost of doing a vacancy survey. In consequence, even where the internet is not widely used, paradoxically, it might be the only reliable source of information to analyse labour demand. Hence, the statistical inference from job portals depends on internet penetration rates; however, even when internet access is relatively low, online job advertisements might be a rich source for analysing important segments of the labour market.

Additionally, as Kureková, Beblavy, and Thum (2014) mentions, it is highly likely that the internet continues to spread across different regions and socio-economics groups, so that reliance on internet-based recruitment methods will

46 According to the Economic Commission for Latin America and the Caribbean (CEPAL 2016, p. 12), around 10% of households in rural areas had access to the internet in 2014, while around 50% of households in urban zones had access to the internet in 2015.

increase over time. In consequence, internet penetration rates limit the statistical inferences that can be drawn from job portal data; however, those limits are becoming less relevant due to technological advances.

4.5.5. Data privacy

Online job vacancy advertisements belong to job portals or to other platforms where employers have decided to make their vacancies public. Provided that job vacancy information is shared and is administrated by a third party, this issue might affect the statistical inferences that can be drawn from those sources. First, the availably of information might change due to modifications on the platforms. As private administrators, job portals might unexpectedly change the number of vacancies or the number and/or kind of variables displayed on their websites, which in turn affects what information is available for researchers, especially when attempting to analyse changes in the economic environment (e.g. number of vacancies, wages, etc.).[47]

Second, job portals can restrict the usage of vacancy information. In most cases, job portals prohibit the storage and usage of job advertisements for commercial purposes; however, for statistical purposes, there does not seem to be any legal restriction. For instance, the Cedefop project "Big Data analysis from online vacancies" has started to collect information from different job portals in Europe. Cedefop has informed these portals that information is going to be collected for statistical purposes, and most of the job portals have not denied access to their data. Nevertheless, as mentioned above, the project has required new statistical legislation to delineate the use of information from job portals and other non-traditional information sources.

Table 4.3 summarises the main advantages and disadvantages of different data sources for the analysis of labour demand. Both traditional (sectoral and household surveys) and non-traditional sources of information (online job portals) have advantages and disadvantages regarding the study of labour demand. Consequently (at this point), non-traditional surveys cannot replace traditional sources of information, although non-traditional sources such as

---

[47] Some websites might adjust the number of variables displayed, such as wages, because potential workers might not apply for the job given previous characteristics of the vacancy.

Big Data might complement and support sectoral or household surveys and vice-versa (see Chapter 9).

Table 4.3. **Advantages and disadvantages of data sources for the analysis of labour demand**

Source	Advantages	Disadvantages
Sectoral surveys	• Guarantee a certain level of data representativeness • Provide (usually macro) indicators of labour demand	• Aggregated data • Time consuming • Relatively expensive • Fixed structure • Less frequent than household surveys
Household surveys	• Guarantee a certain level of data representativeness • Provide (aggregated occupational or skills) indicators about the labour force • Generally available as long-term time series	• Aggregated data • Time consuming • Relatively expensive • Fixed structure • Information from the labour supply
Job portals	• High volume of data • Information in real time • Inexpensive • Disaggregation level • Detailed information • Useful for different purposes (e.g. to estimate vacancy levels, to identify skills and other job requirements, etc.)	• Data quality issues • Job postings are not necessarily real jobs • There is no a priori guarantee of a certain level of data representativeness • Depends on internet penetration rates

Source: Author's elaboration.

4.6. Big Data in the Colombian context

As mentioned before, in some contexts, Big Data sources might be the only ones available to analyse different labour market topics (Kureková, Beblavy, and Thum, 2014). Specifically, in Latin American countries like Colombia, the use of information from online job portals can provide valuable first insights about "skill-shortage vacancies."

Therefore, given the potential for vacancy analysis in Colombia and the high expectations generated by this topic, in order to understand the potential scope of these data sources it is first necessary to answer the following questions:

1) How and to what extent could a web-based system for monitoring skills and skill mismatches using job portals and household surveys be developed for Colombia? Specifically, 2) how can job portal data be used to inform policy recommendations, primarily to address two of the major labour market problems in Colombia, which are its high unemployment and informality rates? And 3) to what extent can these sources be used together (information from job portals [unsatisfied demand] and national household surveys [labour supply]) to provide insights into skill mismatches (skill shortage) in a developing economy?

By answering the above questions, this study contributes to our current knowledge of the advantages and limitations of novel sources of information, which attempt to address public policy issues and/or academic research problems. It provides a methodological and analytical model for countries with scarce information regarding occupations and skills in the labour market by considering possible limitations and biases surrounding vacancy data. It also provides an analysis of the labour market in terms of occupations and skills. Importantly, this research is useful to institutions to match disadvantaged workers (especially unemployed and informal workers) to jobs that they have the potential capabilities to fill, or could be used to help employees develop certain skills, which might not be easily transferable through the formal educational system or programs such as VET (Kureková, Beblavy, and Thum, 2014).

As previously mentioned, the most important ongoing project similar to this book is the “Big Data analysis from online vacancies” project conducted by the Cedefop. So far, this project has been focused on analysing skills and job requirements in Europe from job portals. A remarkable task given the necessity to capture and analyse online sources in more than 24 official EU languages, since April 2018 (Cedefop 2019). However, as summarised in Table 4.4, this study is distinct from the Cedefop project in eight aspects.

Table 4.4. **The main differences between the Cedefop and the Colombian vacancy projects**

Source	Cedefop	Colombian vacancies
Region	European Union	Colombia
Theoretical framework regarding labour market mismatches and the potential usefulness of job portals to tackle skill mismatches	The project is in a stage where vacancy data have begun to be downloaded and processed (exploration stage). It has not been exhaustively discussed and tested to determine the usefulness of job portals to tackle skill mismatches.	It provides a theoretical framework and concepts that highlight the benefits of analysing information from job portals for tacking skill mismatches.
Extraction of information	Job title, skill, and sector variables are collected and processed.	This study considers and proposes various methods to collect and process a wider number of variables, such as job title, labour experience, educational requirements, (imputed and non-imputed) wage, and skills, among others.
Methods to classify job titles into occupations and to identify skills	Machine learning algorithms and the use of a European skills dictionary.	It proposes new mixed methods to properly classify job titles into occupations and to identify skills for a country that does not have national skills dictionaries.
Analysis of variables such as educational requirements, wages, and sector, among others	The project (so far) is focused on describing the most demanded skills and occupations.	This study uses variables, such as occupations, skills, wages, educational requirements, etc., to exhaustively validate and analyse vacancy data.
Period of analysis	April 2018 – ongoing	January 2016 – ongoing
Framework to test the validity and consistency of job portal data	The consistency of the results has not been tested yet.	The vacancy database is exhaustively tested. In fact, a framework is suggested to evaluate its representativeness for each occupation at a different level of disaggregation.
Combination of job portal and household survey data to determine skill shortages	The vacancy data have been used to provide a preliminary overview of demanded occupations and skills.	It provides a descriptive and detailed analysis of occupations, skills, educational requirements, wages, among others. Vacancy data are combined with household data to monitor skill shortages.

Source: Author's elaboration.

4.7. Conclusion

Technological changes have facilitated the generation and storage of large amounts of information at a low cost in terms of time and money. Together with an increased volume of information, a set of different techniques has been developed to process and analyse the massive information generated and available for research and analysis. This large amount of information and the techniques to manage this kind of data have been named "Big Data." As the name suggests, this term refers to a relatively high volume of data; nevertheless, this is not the only characteristic of "Big Data;" indeed, the most common three properties assigned to this term, as described in Section 4.2, are volume, variety, and velocity (Laney 2001).

The Big Data phenomenon has attracted the attention of private and public companies, as well as researchers (among others), because Big Data might provide relevant information for the analysis of individual behaviour, especially in contexts where previously there was a lack of data. The labour market is one of these scenarios where information was traditionally limited or relatively absent, especially for the analysis of labour demand requirements. Collecting information related to labour demand by traditional methods (e.g. surveys) is relatively costly in time and monetary terms. Moreover, even in cases where there is information about labour demand, this information might not be disaggregated (or well-designed) enough to analyse employer requirements. This absence of information and, hence, labour demand analysis is one of the main obstacles to tackle possible skill mismatches. Individuals and training providers unaware of employer requirements might offer skills that are not required by the labour demand.

Consequently, Big Data—specifically job portals—might provide valuable information in real time and at a low cost for the analysis of labour demand, contributing thus to the identification of skill shortages. Compared with traditional sources of information, such as sectoral or household surveys, job portals 1) provide labour market information in a short period of time (real time); 2) enable a relatively inexpensive collection of job portal data; 3) provide a high volume of detailed information, and, hence, 4) allow their data to be disaggregated to skills and occupational levels. Given these advantages and

the potential use of information from job portals, there has been an increasing interest from researchers and policymakers to utilise online job advertisements.

However, little attention has been paid to the possible limitations and biases of the information provided by job portals, and how these issues might affect labour demand analysis. As a source of information, job portal data have the following limitations: 1) data quality; 2) job postings are not necessarily real jobs; 3) data representativeness; 4) internet penetration rates, and 5) data privacy. This chapter has discussed the need for labour demand information that job portals might fill. However, before making any statistical inferences for these sources of information, first it is necessary to know as much as possible about the biases and limitations of their data. Consequently, since Big Data have considerable limitations, as is the case of household or sectoral surveys, it is necessary to evaluate the scope of these sources of information.

At this point, Big Data is a complement rather than a substitute for traditional data collection, such as household and employer surveys, among others. Yet, in a context where information is scarce, Big Data might be the only "reliable" source available for labour demand analysis. This is the case for Colombia (and Latin America), where there are high complaint rates about the quality of workers by companies, and there is not enough labour demand information to address worker skills according to employer requirements. Consequently, the next chapters present a methodology to collect and analyse labour demand information considering possible information biases.





5. Methodology

5.1. Introduction

The analysis of labour demand information is a relevant factor to improve people's skills according to employer requirements. As mentioned by the OECD (2017b), the capacity of countries to improve and adjust their labour supply according to labour demand for skills determines different labour outcomes such as productivity and economic growth, among others, and in the context of this book, unemployment, informality, etc. However, as discussed in Chapter 4, this capacity to analyse labour demand, in most countries, has been hampered by a lack of information about employer requirements.

Recently, online job portals have caught the attention of researchers and policymakers insofar as they might fill the labour demand information gap (Kureková, Beblavy, and Thum, 2014; Reimsbach-Kounatze 2015). These job portals contain a large number of job advertisements, which are accessible to anyone interested in vacancies and employer requirements. Despite this information being publicly available, the analysis of labour demand using job portals is challenging. First, there are large numbers of job advertisements available online, dispersed over different websites; consequently, there is no one consolidated database to use to analyse labour demand information. Second, each job portal manages information according to their own criteria. For instance, some websites might use the term "wage" while others use "salary," or some websites might show remuneration information with numbers while others display them with words or ranges (e.g. monthly £2,000 or two thousand pounds per month, or between £1,750–£2,250 monthly). Moreover, relevant information such as job titles or demanded skills are not categorised to facilitate labour demand analysis.

For the reasons mentioned above, vacancy information is not organised, categorised, and consolidated in a database for statistical purposes. Thus, it is necessary to develop a robust methodology that collects, organises, categorises,

and analyses labour demand using job portals. This chapter proposes and explains each of these methodological steps. The second section of this chapter describes what information is available from Colombian job portals. The third section analyses the most important and reliable job portals to investigate how to conduct a proper labour demand analysis. Given that there is a huge amount of vacancies available online and, consequently, the manual collection of labour demand information is virtually impossible, the fourth section describes web scraping techniques that can be used to automatically collect online job advertisements. The fifth section explains the organisation (homogenisation) of different information from job portals into a single database once the information is collected. Specifically, it explains how programmed algorithms search the information of each vacancy for patterns to build education, experience, localisation, and wage variables from the text. However, not all the variables in the vacancy database can be built using the same method (looking for textual patterns in job advertisements). For instance, to build a variable such as "company sector," it is necessary to implement other and more complex text mining techniques; thus, the last section of this chapter shows how it is possible to identify the sector where the employer belongs.

5.2. Measurement of the labour demand: Job vacancies

As mentioned in more detail in Chapter 2, a job vacancy can be understood as a vacant position within a company that is trying to fill it. Companies recruit potential workers in diverse ways to fill their vacancies. Likewise, as discussed in Chapter 4, job portals provide companies with an informatics platform to make public the number and characteristics of available job positions over a certain period. Even though job portals are not the only channel where companies advertise their vacancies (for instance, occupations related to IT tend to be overrepresented; see Chapter 4), they might capture a large share of the net and replacement labour demand behaviour.

Table 5.1 shows the most important job portals in terms of data traffic (number of visitors) available in Colombia (Alexa 2017).[48] For instance, "https://www.jobportal_a.com.co/"[49] is the 37th most visited web page in Colombia, while "https://www.jobportal_b.com/" is the 89th. Additionally, Column 3 in Table 5.1 shows the number of job advertisements available for each job portal in October 2017.

Table 5.1. **Average number of job advertisements and traffic ranking for selective Colombian job portals**

Colombia	Alexa Rank	Number of job advertisements
https://www.jobportal_a.com.co/	37	115,723
https://www.jobportal_b.com/	89	62,732
https://www.jobportal_c.gov.co/	199	263,621
https://www.jobportal_d.com.co/	1,015	172,440
https://www.jobportal_e.com.co/	2,280	20,143
https://www.jobportal_f.com.co/	3,683	46,853

Sources: https://www.alexa.com and job portals.

A job advertisement is understood as a text on a job portal that shows relevant information about a job vacancy (Swier 2016), and a single job advertisement can contain one or more job vacancies (i.e. mass recruitment). Consequently, the first thing to note from Table 5.1 is that there is a large number of job advertisements and job vacancies on each website.[50] This amount of data makes the manual collection of information a task that would require many working hours and/or a large number of people employed in a monotonous task, that is, copying the information and pasting it in a database thousands of times.

48 Alexa Internet, Inc., is a wholly owned subsidiary of Amazon.com, which calculates and ranks the data traffic of a website based on the browsing behaviour of the internet users of each country.

49 This book anonymised the job portals (removed their names and web page address) in order to protect their identity and to avoid promoting a particular job portal.

50 In Colombia, companies advertise their vacancies on different websites and, depending on the job portal, the cost of promoting a vacancy varies between £24 and £26.

Each job portal shows a list of available vacancies. Nevertheless, each website organises and shows its data according to its own criteria (see Appendix A, Figures A.1 and A.2). Table 5.2 (below) summarises the differences between two job advertisements within the same job portal. Even though this website presents almost the same information about the two vacancies, the localisation and categorisation of these variables (such as experience and wages, among others) might vary according to the website design and the information provided by the employer or recruitment agency. Moreover, some job advertisements on the same website might contain more or less information than the example listed in Table 5.2. Consequently, a job portal is a semi-structured source of vacancy information. This feature makes it difficult to automatically collect data from these website sources. Thus, an algorithm that collects this information needs to recognise differences between advertisements and organise the information in order to properly construct/calculate totals for the database of net and replacement labour demand (hereinafter labour demand database; see Chapter 8).

Table 5.2. **Job advertisement structure comparison within the same job portal**

Variables	Panel A: First job advertisement				Panel B: Second job advertisement			
	Box A	Box B	Box C	Box D	Box A	Box B	Box C	Box D
Job title	X			X	X			X
Experience	X					X	X	
Wage	X			X		X		X
Location	X			X	X	X		X
Publication date	X				X			
Company name								X
Description		X				X		
Number of jobs		X				X		
Educational requirement			X				X	
Type of contract				X				
Workday				X				
Required age							X	

Source: Jobportal_a.

Differences between job announcements also arise when comparing two different websites. For instance, Figure 5.1 compares two job advertisements: one posted on Jobportal_a (Panel A) and the other posted on Jobportal_c (Panel B). Both advertisements require an "accountant" (see Box A, Panel A and Panel B) (note that this is not the same vacancy posted on different websites); nevertheless, the information is displayed in a different way. For Jobportal_a, information about job requirements (such as education, experience, etc.) and job characteristics (such as wage, type of contract, etc.) are shown in Box C (at the bottom of Panel A) and Box D (on the right of Panel A). In contrast, Jobportal_c displays information about job requirements and job characteristics together in Box B (on the left of Panel B).

Additionally, variables such as wages or experience might be categorised in different ways. On Jobportal_a, wages are expressed in numbers (in this case, 1,500,000 Colombian pesos monthly), and the experience requirement is expressed in years. In contrast, for Jobportal_c, the wage variable is expressed in ranges based on the current legal minimum wage[51] and the experience variable is shown in terms of months.

Even though these formatting and structural differences might be regarded as superficial to the human eye, they represent a challenge for the automatic collection of labour demand information. First, structural differences between job portals correspond to differences in how each website was programmed. Specifically, websites can be programmed in different programming languages —such as HTML (Hypertext Markup Language), Javascript, PHP (Hypertext Preprocessor), ASP (Active Server Pages), etc.—; additionally, these languages can be integrated (e.g. an HTML code might contain a JavaScript code). Each of these programming languages possesses its own structure and functions (see Appendix A, Figure A.3).[52]

---

51 In Colombia, every year, the national government decrees the minimum remuneration for a full-time job (current legal monthly minimum wage). For the year 2018, the minimum wage was 781,242 Colombian pesos (around £196) per month.

52 For instance, in HTML, information is delimited by tags, such as "<img/>," "<a>," etc., while information in JavaScript language uses syntax such as "<script type="text/javascript">" "</script>."

Figure 5.1. **Job advertisement comparison between job portals**



Panel A: Jobportal_a[53]

53 Box A stands for: Accountant. Wage 1,500,000 pesos (monthly). City: Bogotá D.C. Department: Bogotá D.C. Posted: Today at 07:04 am; Box B: Company's name: CEYCO Ingeniería S.A.S. Description: Accountant is required, with at least five years of general experience. Contract of service: Part-time. Accounting software: Word office. Date of hire: 03/07/2018. The number of jobs: 1. Box C: Requirements. Minimum undergraduate certificate. No travel is required. Five years of work experience. Age: 20 to 50 years old. Box D: Job summary: Accountant. Company's name: CEYCO Ingeniería S.A.S. Localisation: Bogotá D.C. Working day: Part-time. Type of contract: Contract of service. Wage: 1,500,000 pesos (monthly).

Panel B: Jobportal_c[54]



Sources: Jobportal_a and Jobportal_c.

This heterogeneity between and within websites makes it difficult to automatically collect information. For each job portal, it is necessary to develop an algorithm that recognises the programming language, the structure, and can extract the relevant information from the website and each job announcement. Thus, in order to collect as much information as possible on labour demand, the first part of my methodology involves the following steps:

- Select the most important vacancy websites in the country.
- Scrape the vacancy websites selected.
- Apply text and data mining techniques to organise the information.

---

[54] Box A stands for: Accountant, Transport sector. Box B: Accountant. Company's name: Asistencias Código Delta LTDA. Wage: from 1 to 2 current legal monthly minimum wage (SMMLV for its acronym in Spanish). Indefinite-term contract. Minimum undergraduate certificate. Twelve months of work experience. City: Bogotá D.C. Expiry date: July 5, 2018. Box C: Accountant. Minimum 1 year of experience in accounting, Financial Reporting Standards (NIF for its acronym in Spanish), national taxes, among others.

5.3. Selecting the most important vacancy websites in the country

As shown above, there exist different websites with relatively high data traffic (high number of visitors) and with a significant volume of job advertisements. However, there are a variety of issues to consider before extracting information from job portals. First, there is a trade-off between the number of job portals and the time/effort required to build a vacancy database: as more portals are considered, an increase in effort (human and computational capabilities) and time investment is needed to program each algorithm for each job portal. Additionally, the structure of websites might change over time and, consequently, algorithms need to be adjusted accordingly, and the effort and time to collect information from the selected websites increases significantly as a result.

Second, when considering a larger number of portals, duplication problems arise (as discussed in Chapter 6 in more detail). Companies or recruitment agencies might post the same vacancy on different job portals. As a consequence, the use of many websites increases the probability of duplication. Even though this problem can be diminished by different techniques (see Chapter 6), the probability of duplication persists and even increases by adding more websites. Yet, if a single job portal is used to build a vacancy database, other issues arise.[55] Results derived from that website might be biased or limited in their representativeness of the overall job market. Therefore, in terms of obtaining a certain level of quantity (representativeness) and quality, the selection of job portals is a critical stage in the building of a vacancy database.

Provided that there exist relevant sources of job vacancy information and computational capabilities, to decrease the possible bias of utilising one source, it is necessary to consider job advertisements from different websites to build a vacancy database. In order to select the job portals that best capture the dynamics of the Colombian labour market, the following criteria were applied: 1) volume (number of advertisements available), 2) website quality (structure and number of variables or granularity of information), and 3) traffic ranking (total number of users). Consequently, the methodology proposed here establishes that the

[55] For instance, a job portal might be focused only on a specific segment of the market (e.g. graduate or IT jobs).

selected job portals must have a relatively high number of vacancies, be well-known (traffic ranking) by people, and have a well-defined website structure.

Regarding the former, as shown in Table 5.1, job portals that seemed to have more vacancy information were Jobportal_c (263,621 job vacancies), Jobportal_d (172,440 job vacancies), and Jobportal_a (115,723 job vacancies). However, the volume of posted information should not be the only element to select the most relevant job portals. First, some job portals might post a job advertisement that was originally posted on other job portals. Such is the case for Jobportal_c and Jobportal_d.[56] Consequently, these kinds of websites do not necessarily contain a greater number of job advertisements.[57]

Moreover, the amount of information (number of advertisements) is not the only factor that matters to select the best job portals and to build a vacancy database. The degree of detailed information provided by each website is another element to be considered in the selection process. The more detailed the information, the better the inputs are to build variables such as skills, wages, education, etc. Thus, the second criterion to select a job portal is the granularity of information provided on these websites. In this sense, except for Jobportal_d, the job portals listed in Table 5.1 show similar variables. Indeed, to post a vacancy on these websites, the employer needs to supply a minimum of information (required fields). With some minor variations, this guarantees that these job portals usually have information regarding the job title, city, wages offered, educational requirements, and the company's name, among others. In contrast, the Jobportal_d website does not have a pre-defined format where employers need to fill in the corresponding information. To post a vacancy on this website, it is only necessary to complete the job title, and employers might or might not provide more detailed information in the vacancy description.

[56] Jobportal_d announced that the website had a total of 172,440 job vacancies available on October 30, 2017. However, when clicking on some vacancy announcements, a new window gave a brief and short description of the vacancy and provided the link where that vacancy was originally posted and where an interested person might find more information regarding the job. Similarly, Jobportal_c announced that the website had a total of 263,621 job vacancies available on October 30, 2017. However, when clicking on some vacancy announcements, a new window redirected the search and opened another website where the vacancy was originally posted (e.g. Jobportal_a).

[57] The magnitude of this issue of redirecting was unknown at this stage of the methodology.

Therefore, considering a job portal like Jobportal_d might increase the number of cases with missing values in the vacancy database.

The third criterion to select job portals is the number of users measured by the website's traffic ranking. The number of users might indicate the "trust" of individuals (companies and job seekers) regarding the information provided on a specific website. Additionally, taking into account the traffic ranking of websites guarantees, to some extent, that the selected sites do not specialise in a specific category of vacancies, such as graduate or IT jobs (see Chapter 7 for more evidence regarding this point). As shown in Table 5.1, Jobportal_a, Jobportal_b, and Jobportal_c are the websites with the highest number of visitors.

Table 5.3 summarises the evaluation of job portals conducted in this section. As evidenced in this table, Jobportal_b and Jobportal_a fulfil all the requirements in the consolidation of the vacancy database. These two job portals host the highest number of job advertisements, the variables in the websites are well-defined, and people frequently visit them (traffic ranking).

Table 5.3. **Evaluation of job portals**

Colombian job portals	Real number of job advertisements	Website quality	Alexa traffic ranking
https://www.jobportal_a.com.co/	115,723	Well-defined variables	37
https://www.jobportal_b.com/	62,732	Well-defined variables	89
https://www.jobportal_c.gov.co/	Unknown at this stage (this website posts job advertisements that were originally posted on other job portals)	Well-defined variables	199
https://www.jobportal_d.com.co/	Unknown at this stage (this website posts job advertisements that were originally posted on other job portals)	Not well-defined variables	1,015
https://www.jobportal_e.com.co/	20,143	Well-defined variables	2,280
https://www.jobportal_f.com.co/	46,853	Well-defined variables	3,683

Source: Author's elaboration.

As mentioned above, some websites such as Jobportal_c and Jobportal_d (sometimes) redirect the search and open another website where the vacancy has been originally posted (e.g. Jobportal_a). This redirection issue makes it difficult to know the exact number of observations that each job portal can provide to the vacancy database. Given this uncertainty, the other criteria provide more clarity on which portals should be selected. On the one hand, Jobportal_c has a well-defined structure and has a relatively high traffic ranking. Moreover, this portal is a governmental platform, and it might post governmental vacancies that are not available in other job portals. Thus, Jobportal_c should be considered for the vacancy database. On the other hand, as also mentioned above, Jobportal_d does not have a well-defined website structure. Moreover, there is a considerable difference between the traffic ranking of the first three job portals (Jobportal_a, Jobportal_b, and Jobportal_c) and Jobportal_d. Thus, this job portal does not fulfil the criteria to be considered for the consolidation of the vacancy database.

Finally, Jobportal_e and Jobportal_f have a lower number of job advertisements and low traffic ranking. Additionally, a manual check showed that Jobportal_e and Jobportal_f are not specialist websites that cover job types not found on the three selected job portals. This evidence suggests that reliable information on the total number of vacancies in Colombia might be concentrated in Jobportal_a, Jobportal_b and Jobportal_c websites (Chapter 7 demonstrates that the job portals selected offer a variety of jobs from low-skilled to high-skilled positions).

Consequently, Table 5.4 offers a summary of the web pages that have been selected to be scraped and analysed after an exploration of job portals based on the three elements mentioned above: they have a relatively high number of job announcements (volume), users (traffic), and are well-defined (quality).

Table 5.4. **Job portals and their main characteristics**

Job portal	Main characteristics
Jobportal_a	It is a widespread private platform in Latin America.[58] In Colombia, this source is third in terms of the number of observations (vacancies) posted, it has a minimum number of requirements fields (semi-organised), and it is the most used job portal in Colombia.
Jobportal_b	It is a private platform that operates in Colombia, Costa Rica, Peru, Guatemala, and Salvador. In Colombia, this source is fourth in terms of the number of observations (Colombian vacancies), it has a minimum number of requirements fields (semi-organised), and it is the second most used job portal in Colombia.
Jobportal_c	It is a platform administrated by the Colombian Government (more specifically by the UAESPE). This source is first in terms of the number of observations (vacancies), it has a minimum number of requirements fields (semi-organised), and it is the third most used job portal in Colombia.

Source: Author's elaboration.

Finally, it is important to note that the quality and quantity of information provided by these sources might change over time. Moreover, platforms that were not taken into account in this occasion or new ones might start to provide valuable information (increasing the number of advertisements, increasing the number of users, etc.) in the future. This dynamic might change the decision about which job portals to consider for the construction of a future vacancy database. Thus, the evaluation of job portals should be a constant process to guarantee that the best sources of information are selected to provide the best possible labour demand information.

5.4. Web scraping

As was previously seen in Section 5.2, the differences between and within job portals require modifications in programming language and codification structure. Hence, in order to obtain and analyse labour demand information in Colombia, I implemented a technique called "web scraping," which consists of a computerised method to automatically collect information from across

58 Indeed, there is a version of this platform for Colombia, Peru, Argentina, Uruguay, Guatemala, Ecuador and El Salvador, Honduras, Venezuela, Nicaragua, Cuba and Costa Rica, Mexico, Chile, Panama, Dominican Republic, Bolivia, Paraguay, and Puerto Rico.

the internet (in this case, from vacancy portals) (Oxford Dictionaries 2017). Broadly speaking, this is attained through different softwares that simulate human web surfing to collect specified parts of public information (job advertisements) from various websites and store them in a database to be further organised and analysed.

Although the information is not adequately organised to identify each variable, websites have labels, headers, nodes, tags, among other markers, within their HTML code that allow the extraction of the most relevant information from the data. Codes in R software were built to make this automatic collection of data possible. With the codes developed in this book, the computer can be programmed to visit each job advertisement announcement, to copy all relevant information related to the description of vacancies, and to paste it in a unique database to be organised and analysed. The codes should be built in such way that the computer recognises the job portal's structures, auto-adjusts the number of vacancies to be scrapped, and automatically subtracts and saves the relevant information, among other processes and rules. Thus, to program the codes, knowledge is required in HTML, CSS, and programming languages such as R (see Appendix A, Figure A.3).

Since each web portal displays vacancies in a semi-structured way, they do not follow a well-defined standard to show the data: the Xpaths[59] change between one website and another. Moreover, so far, there is not an automatic way to determine which Xpaths contain relevant vacancy information. As a consequence, the selection of Xpaths needs to be done manually for each website. This selection process requires a certain knowledge of HTML programming language to select the information correctly. Given the difference in the HTML structure from one website compared with another, it is necessary to create a different code for each web portal in order to download relevant vacancy information. In consequence, for this book, this method required the construction of three different codes: one for Jobportal_a, one for Jobportal_b, and the third one for Jobportal_c.[60]

---

59 An expression used to identify nodes in websites.

60 The scraping of each website requires different packages and software. While scraping websites such as Jobportal_a and Jobportal_c does not require sending security credentials (e.g. a login via a user account) to have access to the information, other websites such as Jobportal_b request a login and other user credentials. This login issue (among other issues)

Once the codes are created, the next step is to run the programs in order to download the corresponding information.[61] Each time the codes are run, the (uncleaned) data are saved in a (local) personal server. Importantly, information downloads should be checked periodically. Job portals might inadvertently change their HTML structure. As a consequence, codes might become outdated and fail to extract vacancy information. In this case, the corresponding codes should be updated according to changes in the website structure. However, if there is a long gap between a significant change in the HTML structure of a website and the update of the corresponding code, this might represent an unrecoverable loss of information over a certain time period.[62]

Therefore, it is critical, first, to periodically review (via visual inspection) that each code is extracting the corresponding information, and, second, to run the codes frequently to avoid significant information loss between one download and the next. For this document, each code was run three times per month to avoid information loss.

5.5. The organisation and homogenisation of information

Once the data are obtained, the next step is to provide a well-defined structure to the semi-structured data collected from vacancy portals. As can be observed

makes it necessary to connect R with a software testing framework such as "Selenium" for scraping other websites. Thus, the codes and computing tools (packages and software) to scrape information from job portals might differ between job portals.

61 The process of downloading data using web scraping from a website such as Jobportal_a can last one day, meaning that the computer visits around 80 announcements per minute to obtain the required information, while extracting information from a website such as Jobportal_b takes around three days. These time differences depend on factors such as the web page response time of each job portal, the maximum number of connections allowed, internet speed, sending user credentials, among other factors.

62 For instance, consider a job portal that has 50 vacancies in October 2017, and the corresponding code fails to obtain that information due to changes in the website. In November 2017, the same job portal has 100 vacancies available, 80 of which are new vacancies, while 20 correspond to vacancies published in the previous month (October). Thus, in November, 30 vacancies that were published in October 2017 are not available any more on the website (the jobs were filled and/or the employer paid to post the vacancy for a short period). Consequently, if the code is updated in November 2017, 30 observations from October and their information would have been lost (if the vacancy links are dropped or unavailable on the website).

in Appendix A, Figure A.4, the localisation (XPath) of a variable might change between job advertisements. XPath changes might cause some columns in the database to be out of line. For instance, a column that should correspond to education might contain information about job experience, and vice versa.

Since the information on online jobs boards is semi-structured, it is necessary to use natural language processing techniques to organise vacancy information. Specifically, it is required to use methods to analyse unstructured data such as word analysis (text mining) in order to obtain unified variables, such as wages, work experience, education level, geographic area, and the skills required by employers.

5.5.1. Education, experience, localisation, among other job characteristics

First, it is necessary to carry out a reading of a set of job advertisements to identify the keywords that employers use to describe the characteristics of job positions (such as experience, type of contract, localisation, and education). Once keywords are identified, an algorithm is written in order to "read" job vacancies, which generates a dummy variable that takes the value of one if a particular pattern can be found in a job advertisement (see Appendix B).

Not all variables can be classified into dummy variables, however, given the multiple values that some variables can take, which is the case for localisation, wage, company name, and occupational variables that can accommodate many values, such as the names of different cities, towns, a salary in numbers or words, etc. For this reason, the implementation of another text mining process is required in order to organise and homogenise this vacancy information.

5.5.2. Wages

Employers might or might not provide wage information in job advertisements. When they provide this information, it can take different forms, e.g. wages might be expressed in numbers or words. Moreover, job portals such as Jobportal_b display wage information according to a minimum and maximum range. For instance, a vacancy might contain the following information regarding the wage

offered: "$1.5 a $2 millones mensuales".[63] Given the diverse forms that wage information might take, a number of steps were followed. First, an algorithm was programmed that searches and extracts wage information (in whatever form it takes) from job advertisements.[64]

Second, once the information was extracted and placed in a single column, it was necessary to apply a homogenisation process. As mentioned above, wage information might be displayed in diverse forms. For those cases where the wage revealed the exact number of Colombian pesos that a worker would receive once hired, no depuration was applied, but where wages were described in words, the words were transformed into their equivalent in numbers. Additionally, when wages were shown in ranges, the average value was selected between the maximum and the minimum range. It is important to note that in the above steps, explicit information about wages was sought, and imputation procedures were not yet implemented (Chapter 6 will discuss the issues regarding missing values and possible ways to handle them).

5.5.3. Company classification

The labour demand for skills is produced by a group of private and public companies that perform different activities and provide goods and services. Depending on those activities as well as on goods and services, companies are classified into sectors. Evidently, the required skills might differ from one industry to another sector. Sectors such as mining tend to ask for people with knowledge in controlling heavy machinery for the exploitation of underground mines, while the information and communication sector tends to require people with knowledge in programming. Moreover, there are some generic skills such as communicating and problem-solving that might be used in different sectors. Thus, the analysis of vacancies by sector might identify which skills or occupations are sector-specific or generic.

Frequently, job portals provide information about the company that is advertising a job position. Part of this information might be useful when identifying

---

63 Around £375-£400 monthly.

64 The usage of the Colombian peso ($) symbol or the word "pesos" (which is the Colombian currency) aided the identification of information regarding wages.

the company's sector. On the one hand, websites, in some instances, have a predefined list of sector categories, so that companies are required to select one category when publishing a vacancy (in some cases more than one category to better describe the company's activities); however, job portals have their own classification criteria to create a list of sector categories, and information between one job portal and another might not be comparable. Moreover, sector categories used by job portals might be highly aggregated. For instance, Jobportal_b has the category "services." This option is quite broad and very different types of companies can be classified under this, and therefore the same, sector.

On the other hand, the job description might also give some information regarding the company's sector. However, similar to the above case, companies might use different categories or words to provide information regarding their economic sector. This difference in phrasing is an issue as it suggests that the categories or words used by job portals or companies do not adequately describe company sectors for economic analysis. Fortunately, in most cases, alongside vacancy details, job portals also display the business name of the company that has posted the vacancy. Additionally, in Colombia, the Single Business Registry (RUES, for its acronym in Spanish) database is also available.[65]

Consequently, it is possible to correlate the vacancy and the RUES databases by using company names as a connector between them. However, there are some challenges when trying to merge two databases using company names: misspelling or additional information might exist in either, or both, of the vacancy and the RUES databases. For instance, in the vacancy database, the company's name might appear as "*Exito,*" while in the RUES database the same company might have been registered as "*Éxito S.A.*" Thus, the names that appear in the vacancy and the RUES databases might be not the same, even when they refer to the same company. This possible difference in names between the two databases might complicate the merging of them. Given this issue, it is necessary to utilise word-based matching methods (better known

[65] The RUES is a database where people register their companies to pay taxes and receive government benefits. In this database, company names are available along with other relevant information such as their International Standard Industrial Classification of All Economic Activities (ISIC) code.

as "fuzzy merge" methods) to merge two or more databases using words or sentences (in this case, company names). Generally, word-based matching techniques are a set of algorithms that compare sentences and match phrases that are above a certain threshold matching score. The higher the matching threshold, the more accurate the results, but it is possible that fewer observations are matched; the lower the matching threshold, the less precise the results will be, but it is probable that more observations are matched.

Because different approaches exist—each with their own advantages and disadvantages—to identify the economic sector for each job announcement, this document implemented a combination of manual coding and word-based matching methods (see Appendix C). It is important to note that the procedures implemented in this book are useful to assign an ISIC code to more than half of the observations in the vacancy database (61%). However, the level of disaggregation (4 digits) of this variable might be limited by word-based matching methods or through the use of keywords. For instance, a construction firm might be categorised as "Construction of utility projects" (4220 ISIC code) by observing keyword construction in the company's name. Although such a company might belong to the civil engineering group (division 42 according to ISIC), at a more disaggregated level, it might belong to the construction of roads and railways (4210) (see Chapter 7 for a more detailed discussion regarding this point).

5.6. Conclusion

Information from job portals has caught the attention of researchers and policymakers insofar as it might help to fill the gap regarding labour demand for skills and, hence, improve skills-matching between workers and employers. Nevertheless, processing and analysing information from job portals in a reliable and consistent statistical way is challenging. This chapter has discussed and proposed different solutions to build a robust database of vacancy information from job portals.

Before collecting information from job portals, what is required is a study of the sources to be considered for data analysis. Not every website provides adequate vacancy information. Some job portals provide repeated and/or false information, while other job portals provide a relatively small number of job

announcements. In the case of Colombia, the evidence suggests that vacancy information is well represented in three job portals (Jobportal_a, Jobportal_b, and Jobportal_c). It is important to notice that this number can vary from country to country, and over time.

Once the job portal sample is selected, the next challenge is the collection of thousands of job announcements, both systematically and efficiently. The manual collection of information is virtually impossible. Thus, so far, web scraping techniques are the best way to obtain labour vacancy information from job portals. However, carrying out web scraping techniques requires an in-depth understanding of programming (such as R and Python) and the architecture of each job portal selected in the sample. Each website has its unique HTML structure. As a consequence, web scraping techniques involve programming a different algorithm that automatically and periodically collects information from each website. Moreover, websites might change over time. Thus, algorithms need to be updated whenever there is a change in the HTML structure of the websites of interest.

The challenges for analysing vacancy data do not end with the collection process. Job portals provide detailed information regarding job announcements; however, organising job vacancy information for statistical analysis requires different approaches. Key variables such as wages and required qualifications, among others, are dispersed throughout job advertisements. Thus, it is necessary to program an algorithm that deals with linguistic issues (such as gendered words in Spanish), reads each job announcement, and creates an indicator variable that takes a value (for example, 1) if a particular pattern emerges on a job advertisement. However, in order to build a variable such as company sector, it is necessary to implement other and more complex text mining techniques, such as word-based matching methods (fuzzy merge), and utilise other databases such as the RUES in Colombia.

Moreover, job portals variables might provide information regarding what occupations (at a detailed level of disaggregation) and skills are demanded at a given point in time. Nevertheless, the implementation of different and more sophisticated techniques and processes is required to deduce and organise skills and occupation information. Thus, the following chapter will describe the methods that can deduce skill and occupational information, among other relevant variables.

6. Extracting More Value from Job Vacancy Information

(Methodology Part 2)

6.1. Introduction

The previous chapter has shown that information from job portals might provide detailed labour demand information such as educational requirements and experience, among other vacancy characteristics. However, what makes job portals a potential and remarkable source of data is that they might provide detailed information in real-time about the skills and occupations demanded by companies. As discussed in Chapter 2, the dynamic between the skills or occupations offered by individuals and the skills or occupations demanded by employers is a relevant factor that has strong implications on outcomes for productivity, wages, job satisfaction, turnover rates, unemployment, etc. (OECD 2016a; Acemoglu and Autor 2011). Indeed, the mismatch between the supply and demand for skills might explain a considerable share of unemployment and informality rates in Colombia (see Chapters 2 and 3). Despite the relevance of this topic, detailed information (from official sources such as ONS) for the analysis of the labour demand for skills is relatively scarce due to methodological issues and the high cost of collecting detailed labour demand information (Chapter 4). Thus, the key task of this chapter is to describe the techniques that can be utilised to extract information on skills and occupations.

As mentioned in Chapter 5, information from job portals is not categorised with statistical analysis in mind. For instance, non-categorised information related to skills and occupations (for the Colombian case) can be found in job descriptions and job titles, respectively. Consequently, this chapter explains the steps required to organise and categorise skills and occupational information from the vacancy database. Section 6.2 of this chapter develops a methodology to identify skill patterns in job vacancy descriptions based on international skill descriptors, such as the ESCO. However, there might be some country-specific skills that are not listed in the ESCO dictionary, or its international skills descriptors might not be updated according to the most current labour demand

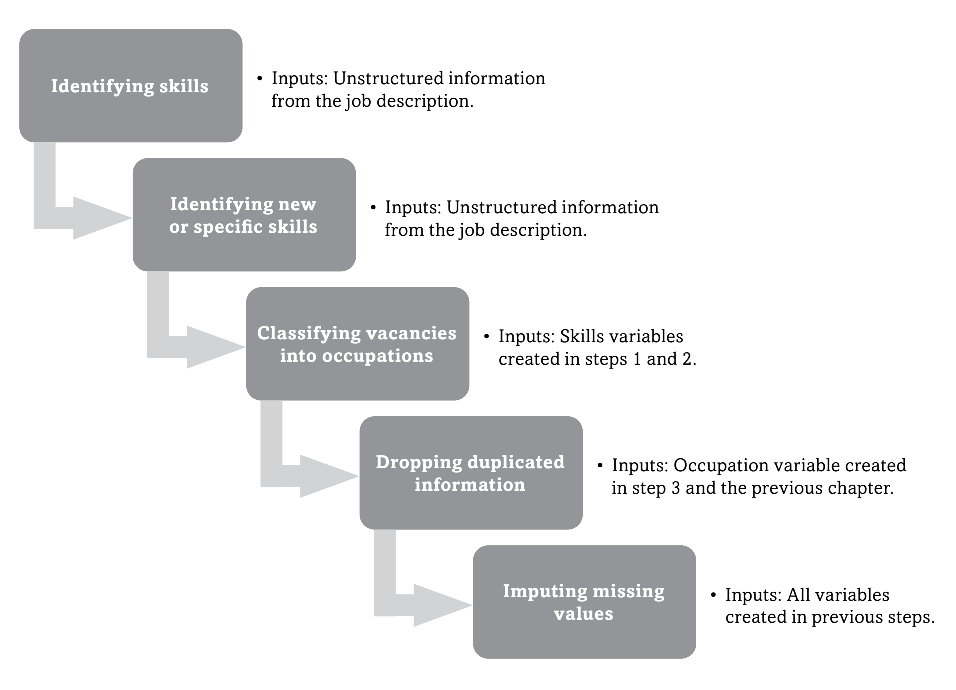
requirements. Therefore, Section 6.3 proposes a methodology to automatically identify country-specific or new skills from information from job portals.

The classification of job titles into occupations is a critical stage for vacancy analysis. Correctly coding the job title variable requires different and advanced data mining techniques. Therefore, Section 6.4 describes and applies techniques such as manual classification, software classifiers, and machine learning to organise job titles into occupational groups. This section also proposes a method that uses unstructured information from job titles and skill requirements (variables created in the previous steps) to identify the occupational groups of hard-coding vacancies. With this last procedure, the vacancy database is completely organised.

Once the vacancy database is organised and categorised into occupational groups, educational requirements, etc., this helps to identify duplication problems at this stage. A job vacancy advertisement might be repeated as an employer might advertise the same vacancy many times on the same job portal or between different job portals. Thus, Section 6.5 deals with duplication issues.

With the vacancy data variables organised and categorised and the duplication problems minimised as much as possible, an imputation process can be conducted for certain variables. As shown in Chapter 5, vacancy data might contain a considerable number of missing values in the variables of interest (e.g. educational requirements and wages offered). This missing information might create biases in the later analysis of labour demand requirements. Thus, Section 6.6 outlines how missing values were imputed for the "educational requirement" and "wage offered" variables by using predictors such as occupation, city, and experience requirements, among others. Finally, Section 6.7 presents consolidated, organised, categorised, cleaned, and imputed data for the analysis of the Colombian labour demand using job portal sources. Next, Figure 6.1 provides a summary of the above described steps that were implemented to organise Colombian vacancy information.

Figure 6.1. **Steps for extracting more value from job vacancy information**



Source: Author's elaboration.

6.2. Identifying skills

As shown in Chapter 5, in most cases, job portals provide abundant information to describe a vacancy. Part of this information is strongly related to the concept of skills, meaning any (measurable) quality that makes a worker more productive in his/her job, which can be improved through training and development (Green 2011) (see Chapter 2 for more discussion on the skill concept). For illustrative purposes, Table 6.1 shows an example of a job description.

Table 6.1. **Job description**

Description[66]
"Importante empresa de sector agroindustrial solicita para su equipo de trabajo analista de calidad. La persona debe tener conocimientos básicos *en sistemas de Gestión (ISO, BPM, Ambiental, SST, RSPO), normativa de calidad, Seguridad Industrial y Gestión Ambiental, buen manejo de Excel y herramientas ofimáticas*. Estudios: Debe tener estudios en *Ingeniera Industrial, Administración, Microbiología, Bacteriología* o estudiante de últimos semestres. Experiencia: mínimo seis meses en cargos o experiencias similares. Funciones: Actualización del *S.G.I de la empresa, recopilar clasificar, registrar, distribuir y archivar la documentación* lo cual incluye correspondencia física y electrónica, redactar documentos diversos para la comunicación interna externa. Salario 836.000 + Prestaciones sociales Lugar de trabajo: Codazzi. interesados enviar hoja de vida actualizada"

Source: Jobportal_a.

As highlighted in Table 6.1, some words or phrases in the job description can be associated with the skill concept. More specifically, words such as "office automation" ("*ofimática*" in Spanish) or "environmental management" ("*gestión ambiental*") can be seen as specific skills required for this vacancy.

Unlike for the study carried out by Lima and Bakhshi (2018), who used pre-defined skills tags to analyse UK job advertisements, for the Colombian case, skills information is not organised under separate variables nor categorised under the same typology. Employers use different words or phrases to describe a skill. Additionally, skills information appears in the job description. Thus, this information needs to be organised to produce informative indicators regarding the labour demand for skills.

As discussed in Chapter 2, there are different ways (typologies or dictionaries) to organise and analyse information regarding skills. Consequently, the first step to organise this kind of information dispersed within vacancy advertisements is to select a dictionary of words or phrases related to skills. Through this

66 English translation: "Important agro-industrial company requires a person with basic knowledge in *management systems (ISO, BPM, environmental, SST, RSPO), quality management standards, industrial safety and environmental management, good Excel management, and Office automation tools*. Studies: Must have studies *in industrial engineering, administration, microbiology, bacteriology* or be a student in the last year of her/his studies. Experience: Minimum of six months in positions or similar experience. Functions: To keep updated the *S.G.I of the company, compile and classify, register, distribute and file documentation*, which includes physical and electronic correspondence, *write diverse documents* for external and internal communication. Salary 836,000 Colombian pesos + Social benefits. Place of work: Codazzi. Interested send resume."

method, it is possible to identify patterns (words or phrases) that are connected to skills in job advertisements. However, Colombia does not have an official dictionary or a list of skills for such a purpose. Consequently, it is necessary to use international references. In this regard, there are different international skill descriptors available, with, perhaps, the most common skills descriptors being used by the O*NET and the ESCO.

As mentioned in Chapter 2, the O*NET is based on the US Standard Occupational Classification (SOC) system. This system contains information on hundreds of standardised and occupation-specific descriptors. Importantly, all these job descriptors are available in the Spanish language; thus, O*NET descriptors can be used to identify skill patterns in Colombian job vacancy advertisements.

The ESCO is a multilingual classification system, so a Spanish version is available for all European skills, competencies, qualifications, and occupations. It is important to note that occupations in the ESCO follow the structure of the International Standard Classification of Occupations (ISCO-08) at the four-digit level, and that the ESCO provides lower levels of disaggregation for each occupation, such as an exhaustive list of 13,485 relevant skills (skills pillar) (European Commission 2017). This list of skills might serve to identify those mentioned in Colombian job advertisements.

Moreover, the ESCO list of skills has an important advantage compared to the O*NET: since the ESCO is mapped following the ISCO-08 structure, the two systems of classification (ESCO and ISCO-08) are compatible. As the ESCO handbook points out: “This is particularly important because most national occupational classifications are currently mapped to ISCO-08” (European Commission 2017, p. 29). Indeed, in 2015, Colombia accepted recommendations made by the International Labour Organization (ILO) to adopt ISCO-08 as official classification.[67] Thus, to obtain results compatible with the official national classification for this book, the ESCO list of skills was employed to identify skills demanded in Colombian job vacancies.

Once the dictionary was selected, the next step was the implementation of text mining techniques to identify the corresponding skills demanded in job advertisements. First, common words in the Spanish language (such as

67 See https://www.dane.gov.co/files/sen/nomenclatura/ciuo/RESOLUCION_1518_2015.pdf.

prepositions, stop words) were removed from the ESCO dictionary and from job descriptions in the vacancy database. Moreover, all letters were transformed to lower case and words were reduced to their grammatical root in both the ESCO dictionary and in the descriptions of the vacancy database. After this, each word or phrase in the skills dictionary was searched for across each job vacancy advertisement. This exploration of words was encoded into unigram variables (n-grams), which are indicator variables. Variables take the value of one if a certain a word or phrase (pattern) of the skills dictionary is found in an advertisement, and zero otherwise.

It is important to notice that each job post does not necessarily contain information regarding skills. There is a considerable share of job vacancies that do not contain skill descriptions. These missing values do not mean that an employer does not require any skills for a particular job, as employers always need workers with a set of skills. However, when publishing a vacancy, employers might not consider it necessary to explicitly write a list of required skills. Consequently, as will be discussed in more detail in the next chapter, unigram variables show the key skills required for a vacancy, but they do not sufficiently identify the complete set of skills needed to perform a job.

Thus, the identification of skills mentioned in job descriptions helps to identify the key skills in demand within the Colombian labour market. Additionally, as shown in Section 6.3, unigram skill variables will serve to identify new or specific skills that are requested in the Colombian labour market and that are not listed in the ESCO dictionary of skills. To have a complete identification of required skills, it is necessary to classify job titles according to an occupational classification (see Section 6.4). Moreover, as will be seen in Subsection 6.4.7, unigram skill variables facilitate assigning occupational codes to the vacancy database.

6.3. Identifying new or specific skills

Although the ESCO dictionary of skills is a complete list for the European labour market, there might be some country-specific skills that are not listed. For instance, Colombian employers might demand different skills compared to Europe. This issue might be the case regarding a specific technology (e.g.

software) that is demanded in Colombia, but not used in Europe. Moreover, as mentioned in Chapter 2, updating dictionaries or occupational classifications might require substantial time, while labour markets rapidly change. This time lapse between changes in the labour demand for skills and the time needed to upgrade skills dictionaries might cause that these dictionaries do not adequately measure what skills are currently in demand.

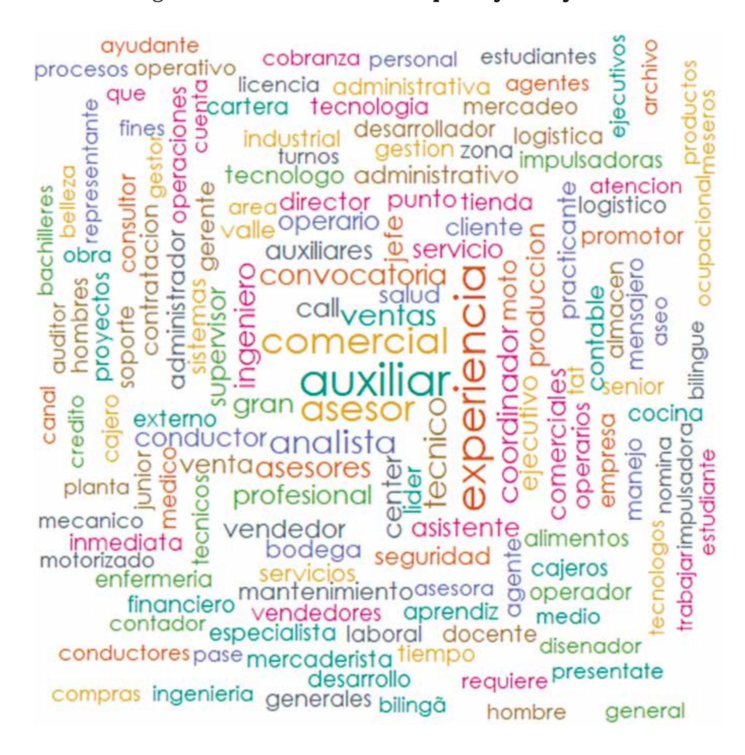
Consequently, to identify new skill patterns from job descriptions, it is necessary to discard information that does not refer to any skill. As in the previous section, common words in the Spanish language (e.g. stop words) were removed from job descriptions. The above technique diminishes a considerable number of words not related to skills; however, a significant number of words might remain that are not relevant to the identification of new skill patterns. As a consequence, a stop words dictionary was created for this study based on the information available in Colombian job vacancies to continue removing non skill-related words. More specifically, column variables from the vacancy database, such as city, wages, type of contract, among others (not related to skills), were used to build a stop words dictionary. The words that appeared in this new dictionary were removed from the description of each vacancy. Nevertheless, several words remained that did not correspond with new skill patterns. For instance, skills identified with the ESCO dictionary remained in the description of the vacancy; consequently, the ESCO skills dictionary was used as a stop words dictionary to remove those skills that were identified previously in Section 6.2. Hence, the words that remain in the description of the vacancy might provide relevant information regarding new and/or specific skills demanded by the Colombian labour market.

It is necessary to note that the words that remain in a job description after applying this method might still contain terms that are not related to skills. For instance, there might appear words related to places or names of people, companies, etc. Moreover, words might appear that refer to other characteristics of the potential worker, such as physical attributes. Consequently, based on the skills definition of this book (see Chapter 2), the final step consists of a visual and manual inspection of the words that remain in the job description to determine which of them are describing new and/or specific skills (Chapter 7 will show a list of new and/or specific skills demanded by the Colombian labour market).

6.4. Classifying vacancies into occupations

One of the most critical variables is "job title" because it summarises the main characteristics of the labour demand and allows classifying jobs into occupations (or skills). Figure 6.2 presents a frequency analysis of the words that appeared in job titles in Colombia in January 2017. According to this figure, the most frequent words appearing in the "job title" variable, and, as a consequence, the most demanded jobs for that time period were: Assistants ("*Auxilar*"), Salespeople ("*Venta*"), Engineer ("*Ingeniero*"), Call centre employees ("*Call center*"), Customer service ("*Cliente*"), Manager ("*Supervisor*"), and Drivers ("*Conductor*"), among others.

Figure 6.2. **Word cloud: Frequency analysis**[68]

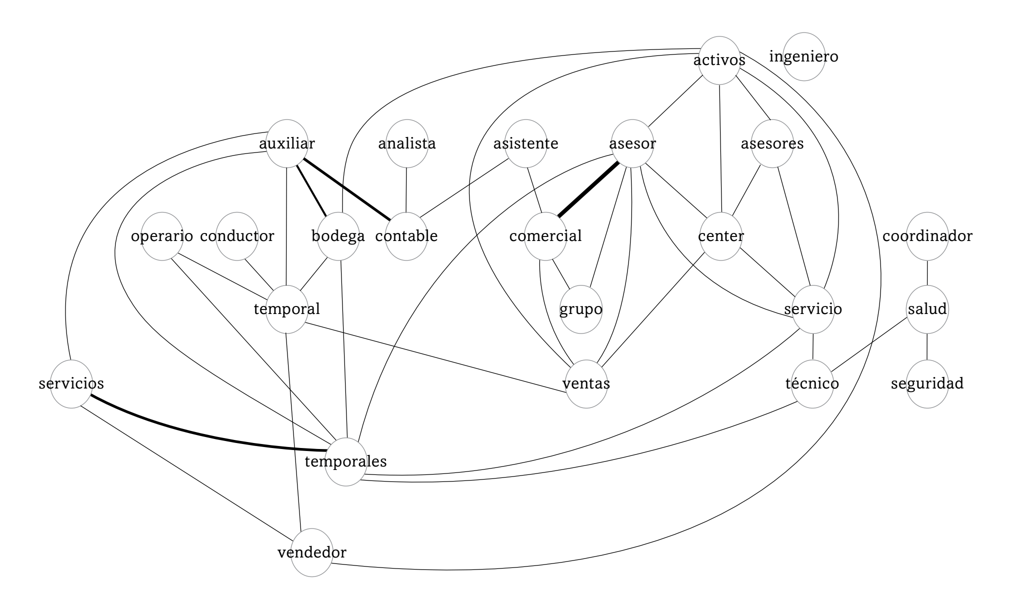


Source: Author's elaboration based on vacancy database, 2017.

[68] The text mining figures are presented in the Spanish language because it is the original language used in the Colombian job portals.

Figure 6.3 shows in more detail the words that are most associated with job titles. A word is related to another word if both words frequently appear together in a job title. Consequently, the thicker the black line in Figure 6.3, the stronger the association is between words. For instance, within the group "Assistants" ("*Auxiliar*"), the most common job title is "Accountant assistants," followed by "Warehouse and services assistants." Within the "Advisor" ("*Asesor*") group, the most frequent occupations are in "Sales, commerce, and customer service."

Figure 6.3. **Word association: Frequency analysis**



Source: Author's elaboration based on vacancy database, 2017.

The above figures are approximations to distinguish the most commonly demanded job titles. However, these figures have many limitations. For instance, they do not identify synonyms. As shown in Figure 6.2 and Figure 6.3, "*Assistant*" and "*Auxiliar*" are considered as different categories, even though they can, on many occasions, refer to the same job category. To avoid these issues and for statistical purposes, it is necessary to use an occupational classification, which is defined as a "tool for organising jobs into a clearly defined set of groups according to the tasks and duties undertaken in the job" (Salazar-Xirinachs 2017).

Regarding job titles, this research seeks to classify all the information available to ISCO-08.[69] However, as Štefánik (2012) points out, there are challenges in transforming job titles into occupation categories because they were created for other purposes. In some cases, there will be more or less information required to classify job titles into occupations. However, such challenges are present in all types of sources, such as household or company surveys, that collect information on occupational titles. Nevertheless, in the case of vacancy data collected from the internet, classifying job titles into occupations might be even more difficult. For instance, alongside job titles might appear the company's name, the city where the vacancy is available, among various words that are not directly related to the job title information. Moreover, as mentioned above, companies might use a variety of different words to describe the same occupation. This issue makes the classification of job titles into occupational codes a challenge.

Given the complexity of classifying job titles into occupations and the importance of this information for researchers, the government, and other institutions, the economic and statistic literature has used three tools to perform the classification process: manual classification, classifiers (Cascot or O*NET API), and machine learning. Manual classification refers to the process where a person or group of people observe job titles. Traditionally, as Gweon et al. (2017) note, assigning occupational codes to texts (job titles) has been a manual task performed by human coders. However, manual classification is a time-consuming and expensive process, especially when handling large databases such as the Colombian vacancy data[70]. Additionally, to guarantee a certain level of coding quality, this manual process would require a professional knowledge regarding occupational classifications and occupation titles. Nevertheless, as Gweon et al. (2017) highlight, manual classifications might provide inconsistent results even with the use of professional coders.

---

[69] As previously mentioned, Colombia accepted the recommendations made by the ILO to adopt ISCO-08 (ILO 2008) as an official classification for jobs.

[70] For instance, the Colombian vacancy data collected for this document in November 2017 consists of around 28,820 job titles (after dropping duplicated titles), and the manual classification of these titles would require a considerable amount of time for a person or a group of people.

More recently, the use of partially or completely automated coding has arisen. Both partial and complete automatic coding significantly reduce coding time. The former term refers to a process where researchers use softwares to set different rules in order to classify certain occupations. For instance, if words such as clerk-bookkeeper or assistant accounts appear in the job title or job description, the set of rules would classify those job titles as "Accounting and bookkeeping clerks" (using ISCO-08). The latter term—completely automated coding—refers to methods such as machine learning. Briefly, these sets of techniques work in the following way: there is an initial stage where the algorithm requires a (representative) training database in which a set of job titles exists, which are already properly classified into occupations (perhaps manually classified). Based on this database, the algorithm "learns" rules of association to code job titles. With this knowledge, the algorithm can predict the most probable occupational code for each job title for new data (Gweon et al. 2017; Lima and Bakhshi 2018).

Moreover, there exist softwares such as Cascot (Computer Assisted Structured Coding Tool[71]) (Jones and Elias 2004) (see Subsection 6.4.3) that allow both partial and/or complete automatisation. This kind of software already contains a set of logic rules. Based on a score of similarity between occupation titles (provided by the occupational classification, e.g. ISCO-08) and job titles (e.g. posted on job portals), the software assigns a corresponding occupational code (which has the highest similarity score). In this way, a list of job titles can be automatically classified. However, complete coding automatisation was still a challenging process at the time when this book was written due to the complexity of categorising occupational titles (Gweon et al. 2017). Besides, algorithms fail to provide a perfect classification for each job title (Belloni et al. 2014).[72]

71 Developed by the Institute for Employment Research (IER) at the University of Warwick.

72 To avoid misclassifications, Jones and Elias (2004) recommend the implementation of both partial and fully automated coding (semi-automatic coding). For instance, in the case of Cascot, the authors suggest automatically classifying all job titles (inputs) and keeping a record of similarity scores. For those job titles where the similarity score is below a minimum threshold, it is necessary to assign a corresponding occupational code manually. In this way, the time spent classifying job titles into occupations will decrease, and a certain level of coding quality will be guaranteed.

Thus, given the availability of several tools to classify occupations and the advantages and disadvantages of each one of them, the next sections will discuss manual coding, cleaning, automatisation, and adapting Cascot.

6.4.1. Manual coding

As pointed out before, manual coding is a time-consuming task. However, as shown in Figure 6.2 and Figure 6.3, there are some job titles that are more frequently mentioned by employers, hence those job positions constitute a considerable share of the vacancy database. Additionally, automatic algorithms might misclassify some job titles, given that automatic methods of classification might fail to classify some job titles that appear with more frequency in the vacancy database. As a consequence, coding quality might be primarily affected by the misclassification of some common job titles.

In order to ensure that the most frequent job titles are adequately classified, a careful manual coding process was carried out for job positions that were more numerous, and, therefore, it was relatively easy to determine their occupational groups. Moreover, as words in the Spanish language are gendered and words might slightly differ in the plural and the singular, the roots (patterns) of the words were used instead of looking for exact combinations of words. For instance, manually classified titles such as "Accountants" were extracted by using the root "*Contador*" instead of "*Contadora*" for a woman or "*Contadores*" in plural. By doing so, a total number of 50 job titles received an occupational code (which corresponds to around 27% of the job advertisements). This information suggests that a considerable share of the Colombian vacancy information is concentrated across relatively few job titles.[73]

6.4.2. Cleaning

As mentioned above, coding quality depends on the tool used and on the quality of input data. However, job titles displayed on job portals sometimes contain extra information (noise) that might affect coding quality. While there

---

[73] At this point, this result neither validates nor invalidates the reliability of data. The Colombian labour market might demand a particular set of occupations (see Chapter 7 for further discussion).

are some group words such as prepositions that might be easy to identify and clean from the data, there are other words that do not belong to a specific group of words that frequently appear in job titles and do not describe a job position.

As shown in Figure 6.2, in the job titles, abundant information is not directly related to the job position (such as company name and working hours). It is common to see words such as “time,” “immediately,” and “required,” among others, in the Colombian vacancy data. The presence of these words might affect the performance of automatic classifiers. In order to assign an occupational code, tools such as Cascot or the ONS Occupation Coding Tool compare the similarity of words in the job title from a job vacancy (or another source of information) with a directory of job titles. The extra information might affect this comparison. For instance, when the input is “Accountants” with a similarity index of 92, Cascot assigns the ISCO code 2411 (“Accountants”)—in a scale of 0 to 100, the higher the number, the higher the degree of certainty that a given code is the correct one. However, when the input is “Accountants immediately,” the similarity index drops to 66.

Thus, before conducting automatic classification processes, the job title variable, which is the primary input to assign an occupational code, was carefully cleaned. First, prepositions, adverbs, nouns, among others, were dropped from data. Second, the variables “city” and “company name” (provided that the structure of the website contained this information) were used to identify all possible locations and employer names that might arise in job title variables. With this process, names that might appear in the job title were dropped. Third, with a visual inspection of the vacancy database and the usage of word clouds, it was possible to identify and drop those words that did not contain information regarding occupation in the job title. After this manual cleaning process, automatic classification tools and techniques were used.

6.4.3. Cascot

The first step in the automatisation process is the usage of Cascot. As mentioned before, this tool was developed by Jones and Elias (2004) at IER. Cascot is designed to assign an (occupational or industrial) code to texts. In the case of occupational classification, Cascot allows the classification of a piece of text (job titles) according to their UK Standard Occupational Classification

(SOC 1990; 2000; 2010). Moreover, since 2014, a multilingual ISCO-08 version of this computer program has been developed for nine languages (Dutch, English, Finnish, French, German, Italian, Portuguese, Slovak, and Spanish). Additionally, in 2016, the software was extended to another five languages (Arabic, Chinese, Hindi, Indonesian, and Russian).

This multilingual capability is one of the most critical characteristics of Cascot. It allows classifying job titles from different languages into occupations following an international standard such as ISCO-08. In order to classify a piece of text into an occupational classification (e.g. ISCO-08), Cascot has a set of rules—such as downgraded words, equivalent word ends, abbreviations, replacement words, word alternatives, etc. (Warwick Institute for Employment Research 2018)—that reveal the best matches between job titles (inputs) and occupational classifications with corresponding similarity scores. Importantly, to set up all the association rules (mentioned above), the IER made partnership arrangements with experts for each country covered for the testing and refining of Cascot (Wageindicator.org 2009).

Moreover, Cascot outputs have been compared with high-quality and manually coded data (Jones and Elias 2004). According to this test, 80% of records that receive a similarity score higher than 40 coincided with the manually coded data. Thus, Cascot offers, to a certain extent, a well-defined directory of job titles with occupational codes and association rules that can be used for coding job titles.

Consequently, one of the main reasons to use Cascot is that it already has a deep and reliable knowledge base, built over years. Indeed, relatively new classification methods such as machine learning should consider and "learn" from the association rules that have been created through years of research using Cascot. Moreover, this tool has a considerable advantage in a context where there does not exist (or at least is not publicly available) a trustworthy pre-processed database with job titles and occupational codes. Machine learning methods need as input a training database (which is data that were previously and correctly classified). Without this training database it is not possible to use machine learning models to assign occupation codes.

Taking the above reasons into account, Cascot was used to classify job titles in the Colombian vacancy database. Following the recommendations of Jones and Elias (2004), Cascot assigned an occupational code to a job title

when the similarity score was greater than 45. This threshold was to re-ensure that Cascot outputs would coincide with the manual coding revision in most cases. By doing so, around 38% of the observations in the vacancy database received an occupational code at the four-digit level.[74] Thus, 35% of the job advertisements required further data management to assign a proper occupational code.

6.4.4. Revisiting manual coding (again)

Provided that 35% of the database was "hard-coded" (not classified by Cascot), it was necessary to conduct another short manual coding process. Here, the same methodology explained in Subsection 6.4.1 was applied. First, a visual inspection of the vacancy database was conducted on data that were not classified by Cascot. Job titles that appeared more frequently in the database were manually assigned an occupational code. Once again, the usage of the roots of the words was necessary to avoid any issue with gendered or plural (singular) forms. This ensured that hard-coded job titles that were more frequent in the vacancy database received a proper occupational code. In total, 50 job titles were manually coded, which corresponds to around 5% of the total number of job advertisements. At this point, approximately 70% of the observations were assigned an occupational code with a relatively high standard level of confidence.

6.4.5. Adaptation of Cascot according to Colombian occupational titles

The ISCO contains a standard list of occupational titles used in the international workplace, which is linked to categories in its classification structure. This list is a key input for Cascot to match occupational codes and job titles. However, as mentioned by the ILO (2008, p. 68): "[occupational titles provided by ILO] might be a good starting point to develop a national index. The national index,

74 A sample of those observations was selected to evaluate the accuracy of the Cascot tool for the Colombian case. According to this manual check, around 94% of the observations had the correct occupational code (ISCO-08) at a four-digit level. Moreover, common mistakes were manually corrected.

however, needs to reflect language as used in survey responses in the country concerned." Even in countries with the same language, job positions might be named differently depending on the national context.[75] Consequently, standard occupational titles provided by the ILO might not cover a considerable share of Colombian job titles, hence Cascot might not assign an occupational code to a high portion of them. Indeed, this issue of context might explain that, at this point, only 38% of the job portal observations were categorised using Cascot.

Moreover, the DANE released an adaptation of the ISCO occupational titles according to the Colombian context in 2015 (DANE 2015). Thus, given that Cascot can be edited, the adjustment of the Colombian occupational titles can complement this tool. Consequently, the following step was updating Cascot to the Colombian context by using the occupational titles utilised in this country. Once this adaptation was ready, the job titles that had not been coded in the previous steps (around 30% of the total number of job advertisements) were processed once again for Cascot with the same specifications mentioned in Subsection 6.4.3. Interestingly, with the adaptation of the tool, around 12% of the total number of advertisements were assigned an occupational code. Thus, by only adapting the Cascot tool using the national occupational titles of Colombia, the portion of job advertisements has considerably increased from 70% to 82%.

However, concerns might arise regarding the accuracy of coding with this adapted version of Cascot. Regarding this concern, it is necessary to highlight that the occupational job titles used to adapt Cascot come from the national statistical department in Colombia and are publicly available. Moreover, the list of Colombian job titles is the product of the joint work of institutions such as the DANE, the Ministry of Education, the Ministry of Labour, and training providers, among others (DANE 2015). Thus, the input "occupational titles" should be similar to job titles in job advertisements.[76]

---

75 For instance, in Colombia, there is a particular job title to define general maintenance and repair workers, which is "*Todero*." This job title cannot be found in countries like Peru or Chile (where Spanish is also the official and most spoken language).

76 A manual check was carried out to determine the accuracy of correctly coded observations. According to this manual check, around 92% of the observations had the correct occupational code (ISCO) at a four-digit level. Moreover, common mistakes were manually corrected.

6.4.6. The English version of Cascot

As a result of the above described manual check, a considerable portion of job titles that were found to lack an occupational code were those written in English. Despite Spanish being the official language of Colombia (among other minority indigenous languages), job titles such as "Customer care analyst," "Data analyst," "Courier," etc., are written in English. Consequently, the English version of Cascot might help to classify some of the job titles in the vacancy database. However, the English version of Cascot assigned an occupational code to a job title if the similarity score was greater than 60. This threshold is set at 60 to avoid any confusion and misclassification with job titles in the Spanish and English Cascot versions. By doing this, 3% of the job titles in the vacancy database received an occupational code.

At this point, 15% of the observations remained without an occupational code. There were three options for classifying the remaining job titles: 1) manual coding, 2) using lower minimum similarity threshold through Cascot, or 3) other techniques such as machine learning. The first method, as explained repeatedly above, is a time-consuming task. Therefore, this option was not considered. At the same time, the second and third options contain various advantages and disadvantages. On one hand, the Cascot similarity threshold could be lowered to classify more job titles (so far, the threshold has been 45). Nevertheless, this might increase the number of misclassified observations.[77] On the other hand, machine learning techniques could serve as a complement to identify occupations. As mentioned previously, machine learning techniques have been implemented during the last research year to assign occupational codes to job titles. Depending on the sophistication of their algorithms and

[77] This option is the most straightforward alternative to assign occupational codes to the remaining observations because it is relatively easy to conduct. Although Jones and Elias (2004) recommend using a minimum threshold of 40, each researcher can reduce this threshold and increase the number of observations with occupational codes. However, this might also increase the number of misclassified observations. The Cascot minimum score threshold was lowered to 30. This minimum threshold was set arbitrarily as a starting point to evaluate Cascot's performance. A sample of observations with a threshold of 30 was taken to assess Cascot's performance. As expected, the accuracy level of automatic coding decreased. Around 39% of the job titles were incorrectly classified. Thus, lowering the Cascot threshold was not an option to classify the remaining job titles.

inputs (training and test databases), these techniques might adequately assign occupational codes to job titles (Bethmann et al. 2014).

6.4.7. Machine learning

The use of machine models that classify job titles into occupation codes has arisen over the last decades. As Gweon et al. (2017) highlight, institutions such as the Australian Bureau of Statistics have favoured this method. In concrete terms, machine learning is a "set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" (Murphy 2012, p. 1).[78] Moreover, as Murphy (2012) points out, classification (Supervised Learning)[79] is perhaps the most commonly used form of machine learning to solve real-world issues. The idea in this method is to classify a "document," for instance, a job title, into one of several classes (C) based on some previously learnt training inputs (X). The computer determines how to classify a document based on both a training dataset and a particular association algorithm. The former refers to a pre-processed dataset with an N number of training examples (usually denoted by D). For the case of job titles, this database is a pre-processed database with job titles assigned with corresponding occupational codes (see Appendix D).

In terms of assigning occupational codes to job titles, the economic and statistic literature has favoured SVM (Support Vector Machines) (Gweon et al. 2017) (see Appendix E). However, as Appendix F demonstrates, 40% of the vacancy job titles were incorrectly classified by using SVM. Therefore, the SVM machine learning algorithm, which only uses job titles, is not an option to classify the remaining observations in the vacancy database.

[78] These machine learning methods have been applied in several fields, such as health and economics, among others (Varian 2014; Zhang and Ma 2012).

[79] Unsupervised and reinforcement learning are other types of machine learning algorithms. However, as the purpose of this subsection is the classification of job titles, this document is focused on Supervised Learning.

6.4.7.1. Nearest neighbour algorithm using job titles

As shown in the previous subsection, the numeric transformation of job titles with the SVM algorithm might serve to assign an occupational code to hard-coding observations. However, the number of job titles classified by SVM is limited, and, consequently, it is necessary to use more advanced techniques to code job titles. In this regard, Gweon et al. (2017) demonstrate that (with some adaptations) the nearest neighbour algorithm might provide better results regarding accuracy than the SVM algorithm. Briefly, the nearest neighbour algorithm takes new record(s) (in this case, n-grams of a job title), maps this (these) new record(s) in the training dataset, and finds the closest observation to this new record based on n-grams of the job titles. Once the nearest neighbour(s) is (are) selected, the algorithm assigns to the new record(s) the class (y) of its (their) closest neighbour(s) (see Appendix G).

6.4.7.2. Machine learning using skills

Conversely, Lima and Bakhshi (2018) proposed an extension of the basic machine learning model for classifying job titles into occupations. For this study, the authors used UK job vacancies published in 2015, collected by Burning Glass Technologies.[80] This company assigns each vacancy one or more of 9,996 tags derived directly from the job advertisement text (the authors did not clarify, though, how and based on what the tags were built). Consequently, instead of using job titles (n-grams) as an input to assign an occupational code to each observation, the authors propose to use a naïve Bayes algorithm that takes as its predictors (x) the skills mentioned in the vacancy advertisement. By doing so, Lima and Bakhshi (2018) demonstrated that a skills-based classifier might improve the coding of jobs titles that are poorly classified.

---

80 Burning Glass Technologies is a company that provides job market analytics.

6.4.7.3. Nearest neighbour algorithm using skills and job titles

Given these advantages and limitations of the more recently proposed algorithms, this document uses an extension of the algorithm proposed by Gweon et al. (2017) by adding the n-grams (input x) information related to skills, as suggested by Lima and Bakhshi (2018). Specifically, it is recommended to complement n-grams (input x) from the job title with the skills mentioned in the job description. Skills information is supposed to be highly correlated with the job title. For instance, for a job position such as "Secretary," it is logical to think that employers will require relatively more skills related to office automation, while for a job position such as "Kitchen helpers," the skill requirements will be relatively more related to food production. Consequently, by considering the skills demanded and the job titles, it is possible to find a more similar training dataset that might improve automatic coding (see Appendix G).

6.4.7.3.1. Application of the extended nearest neighbour algorithm to the vacancy database

As mentioned in Subsection 6.4.6, 15% of the job titles remained uncoded at this stage through manual and Cascot procedures. Consequently, the final step to classify the remaining job titles was conducted using the extended nearest neighbour algorithm, explained in Appendix G (Tables G.5 and G.6). However, as pointed out in Section 6.2, unlike the study of Lima and Bakhshi (2018), where the authors had at their disposal pre-defined skill tags to use as inputs for the machine learning model, for the Colombian case, skills information (which is the key input required to implement an extension of the nearest neighbour algorithm) is not organised into separate variables, nor categorised under the same typology. Thus, this book uses the n-gram skill variables created in Section 6.2 as an input for the algorithm proposed here.

Specifically, the 1,910,000 observations (85% of the vacancy database) coded from Subsection 6.4.1 to 6.4.6 were used as input to train and test the extended nearest neighbour algorithm. Each of those observations has the corresponding 4-digit-level ISCO codes and the n-gram skill variables identified in this book. Moreover, this input database was divided into two: training and test database. Following Dobbin and Simon (2011), the training dataset

is composed by 1,273,333 (two-third) observations (randomly assigned) from the input database, while the test database is composed of the remaining one-third of the input data. The computer determines how to classify the job titles by executing the extended nearest neighbour algorithm with the training database. Once the computer had learnt the association rules, the algorithm was executed in the test database. The predicted results were compared with real ISCO codes in the test dataset. The comparison showed that the extended nearest neighbour algorithm correctly classified 92% of the test dataset. Thus, the algorithm showed a high accuracy prediction level.

By doing so, this book uses an algorithm (nearest neighbour) with a proved high accuracy level for categorising job titles. Moreover, using skill n-grams based on the ESCO dictionary shows that the description might increase the accuracy level and the number of job titles coded without the need for pre-defined skill tags (see Appendix G for a comparison between the accuracy level of these algorithms). With this method, 10% of the job titles were coded. Consequently, at this point, 95% of the job titles in the vacancy database have received an occupational code.

Despite machine learning methods and classifiers such as Cascot significantly reducing the time spent on coding, at the time of writing this book, it is still necessary to use manual coding for those job titles that remain unclassified. Consequently, 50 job titles were coded manually. Thus, through automatic and manual processes, 96% of the job titles were coded according to ISCO (4-digit level).[81]

[81] Importantly, a considerable percentage of non-classification might be explained by the absence of key information in the job title variable. The most frequent words in those job titles without an occupational code do not provide information regarding the job position. For instance, a regular word is "*Bachilleres*" (which in English means "undergraduate"). Clearly, with only these kinds of words in the job title, it is not possible to identify their requirements through automatic or manual means. One reasonable alternative to overcome this issue is to take into account the job description. Perhaps information about the job position is in the description rather than in the job title. Thus, processing and identifying specific patterns in job descriptions might increase the number of observations with an occupational code. This further development will be part of a future work.

6.5. Deduplication

Along with the categorisation challenges shown above, there is another important issue to consider, which is the possibility of duplicated information. As data are collected from different websites, some job advertisements can appear on more than one job board, or even several times on the same job board (Chapter 4). This issue can result in a significant over-counting of job advertisements and might affect the results when data are analysed. For those reasons, before data analysis, it is necessary to apply a measure to identify which vacancies are duplicated to discard all but one of them. This process is known as "deduplication" (Carnevale, Jayasundera, and Repnikov 2014).

One option is to drop those vacancies that have the same job title, level of education, city, sector, date published, wages, etc. However, this string-based approach is not enough to completely solve the duplication problem, e.g. an employer can post a vacancy with the job title "Taxi Driver" on a website, and another website can write "Taxis Driver" for the same vacancy. With the method described above, this vacancy would count as a different one. Therefore, it is necessary to develop or adopt a measure of "similarity" to decide the probability with which an observation is duplicated. In this regard, Gweon et al. (2017) have shown that n-gram-based methods for dropping duplication in job titles are preferable than string-based methods. As mentioned in Section 6.2, n-grams are a set of indicator variables based on text patterns. The variables take the value of one in the presence of specific patterns.

Consequently, n-gram-based methods are not sensitive to minor changes in string variables (such as the job title). Thus, following Gweon et al. (2017), an n-gram based method was applied to drop the maximum number of observations duplicated. More specifically, a duplicated job advertisement was discarded if the values of the previously created dummy variables (such as experience, educational requirements, type of contract, localisation, and wages) were the same as in other job advertisements, including their ISIC (Chapter 5) and ISCO codes (Section 6.4), the publication date, and the number of job positions required. By doing so, around 26% of the observations were discarded.

6.6. Imputing missing values

Provided that the information comes from websites and employers who might not provide a full description of the vacancy, there are variables with missing values. For instance, despite the text mining techniques explored in Chapter 5, around 30% of the observations in the "wage" variable have missing values. As the presence of missing values can create biases in the analysis (Little and Rubin 2014), it is essential to implement imputation techniques to analyse full data vacancy information.

In this regard, Carnevale, Jayasundera, and Repnikov (2014), with hot-deck and cold-deck methods, imputed missing educational requirements in job advertisement data using a combination of the education distribution of the vacancy data (no missing values) and the education distribution of employment (from the American Community Survey, ACS). With such a method, they demonstrated that it is possible to use the whole vacancy database to test whether the information contained in it is representative of different education levels.

Given the relative importance of the analysis of labour demand for skills and the considerable presence of missing values in the data, for this document, an imputation procedure is conducted for the wage and educational variables.

6.6.1. Imputing educational requirements

For the Colombian case, 20% of the observations in the educational requirement variable contain missing values. These missing values do not mean that for those vacancies Colombian employers do not have any educational requirements. Employers might forget to mention educational requirements, or information regarding education might be implicit in other variables (such as the job title). Indeed, in most of the job titles in the vacancy database, the educational requirements are implicit. For instance, job titles, such as lawyer, economist, and psychologist, among others, implicitly reveal that employers require a worker with at least university education.

Consequently, to impute missing values, a hot-deck imputation was conducted as proposed by Carnevale, Jayasundera, and Repnikov (2014). Specifically, through this method, an observation with a missing value in a particular variable receives a value, which is randomly selected from a sample ("deck")

of non-missing records that have some characteristics ("deck variables") in common with the observation with the missing value. For instance, for the Colombian case, an observation with a missing value in "educational requirements" receives a value from an observation that is randomly selected from a sample of records, which have the same characteristics in common, such as the same occupation. Consequently, as a first step, it is necessary to define what characteristics define the sample of donors ("deck") for an observation with a missing value.

Within a vacancy, the variables occupation, city, and year were considered as characteristics that defined the sample of donors. By using these three variables, it is possible to establish a proper sample of donors for observations with missing values about educational requirements. The occupational variable (at a 4-digit level) guarantees that both the donors and the missing observation(s) contain similar skills and tasks. Indeed, the occupational variable is the most important factor of the imputation process because, as mentioned above, occupation (job title) is a concept strongly related to educational requirements.

Additionally, examining the city (where the vacancies were posted) controls for possible differences in educational requirements from one place to another (e.g. a city to a town). The year of the vacancy controls for the fact that educational requirements change over time. As Spitz-Oener (2006) notes, to perform a particular occupation today involves greater complexities than at the end of the 1970s. For instance, in the past, it was enough to have a high school certificate to apply for a job as a secretary; now, for the same job title, it is necessary to have a higher educational level given technological changes, among other factors. Besides these, no other characteristics in the vacancy database were taken into account due to the high presence of missing values in those variables (e.g. wages).

Thus, an observation with a missing value in "educational requirements" receives a value from another observation if, and only if, that record was offered in the same city and year and has the same occupational category. It is important to note that this book did not implement the cold-deck method. In contrast with the hot-deck method, cold-deck imputation picks donors from another database; for instance, from household surveys. This book does not use the cold-deck method for the following reasons. First, the frequency of missing values in educational requirements is not as high compared to the

study by Carnevale, Jayasundera, and Repnikov (2014), where roughly 50% of the vacancies have a missing value in their educational requirements. Thus, for the Colombian case, there is enough information with no missing value (80%) to impute the remaining missing values.

Second, and more importantly, the cold-deck method proposed by Carnevale, Jayasundera, and Repnikov (2014) uses the American Community Survey (ACS) (which is a survey on labour supply) to impute missing values in the job vacancy data. However, as will be discussed in more detail in Chapter 7, missing vacancy values based on a household (supply) survey might be problematic due to the distribution of educational requirements (among other characteristics) that might differ between labour demand and labour supply. Moreover, part of this book seeks to test whether the vacancy database shows consistent patterns compared to official statistics such as household surveys. Consequently, the implementation of a cold-deck method with a household survey imposes on the vacancy database a distribution of educational requirements related to labour supply, and thus any comparison in terms of educational level between labour demand and supply might be affected by this cold-deck imputation process.

6.6.2. Imputing the wage variable

Finally, given the importance of wages for labour demand analysis and the presence of a missing value for this variable in the Colombian vacancy database (around 30% of total observations), an imputation procedure was conducted. Traditionally, imputation methods involve linear or logistic regressions; however, as Varian (2014) mentions, when a large amount of data are available, better methods to impute variables can be applied, such as the LASSO regression ("Least Absolute Shrinkage and Selection Operator"). Unlike linear models, the LASSO model penalises predictors that do not have relevant information and might increase the error term (e) for predicting an output (y)—in this case, the missing values for the wage variable (Varian 2014). In other words, the LASSO model selects and drops those predictors (variables) that do not contribute to wage prediction.

The occupation variable might be comprised of 40 different values (sub-major ISCO groups), for instance, which means that for the LASSO model, those values in the occupation variable are transformed into 40 dummy variables.

Specifically, to impute the wage variable (y) in the vacancy database, the following was conducted:

$$y = \beta_i\, Occupation_i\, \chi_{\{i=1\ldots40\}} + \beta_i\, Department_i\, \chi_{\{i=1\ldots32\}} + \beta_i\, Quarter_i\, \chi_{\{i=1\ldots4\}}$$
$$+ \beta_i\, Education_i\, \chi_{\{i=1\ldots8\}} + \beta_i\, Workday\, \chi_{\{i=1\ldots3\}} + \beta_i\, TypeContract_i\, \chi_{\{i=1\ldots4\}} + \varepsilon$$

Where *y* is the wage variable, "*Occupation*" denotes the set of dummy variables that identify occupation (ISCO two-digit level, 33 subgroups);[82] "*Department*" represents the set of dummy variables that identify the department where the vacancy is available (there are 32 departments in Colombia); "*Quarter*" denotes dummy variables that indicate the quarter of the year when the vacancy was downloaded; "*Education*" represents a set of dummy variables that indicate educational requirements (six categories[83]); "*Workday*" and "*TypeContract*" are sets of dummies variables indicating the workday (three categories) and the type of contract (four categories) offered by employers (all of these categories will be explained in more detail in Table 6.2).[84]

6.7. Vacancy data structure

Figure 6.4 summarises the steps carried out and the amount of information processed to consolidate the vacancy database for Colombia.

82 The occupation variable was grouped at a two-digit level to avoid oversaturation and due to computational limitations.

83 Due to frequency issues, the categories of specialisation, master's degree, and doctoral degree were grouped in one category: "Postgraduate."

84 The variable sector was not included in the imputation model due to the high frequency of missing data.

Figure 6.4. **Summary of steps carried out to obtain the Colombian vacancy database**

Selection of the most important vacancy websites in the country: Three job portals (Jobportal_a, Jobportal_b, and Jobportal_c) fulfill the volume, quality and web traffic criteria to conform the Colombian vacancy database.

The information from each website was scraped every ten days for three years. The total number of monthly collected vacancies was around 38,200 for Jobportal_a, 25,500 for Jobportal_b, and 22,000 for Jobportal_c. Consequently, a total number of 3,037,868 vacancies were collected during the study period.

Organisation and homogenisation of information, such as education, experience, localisation, occupations, among other job characteristics.

Deduplication: 26% (789,845) of the observations were dropped because they had duplicated values. A total number of 2,248,023 (not duplicated) vacancies were collected during the study period.

Imputing missing values:
Educational requirements: 20% of the observations in the educational requirement variable were imputed using a hot-deck imputation method.
Wage variable: 30% of the observations in the wage variable were imputed using a LASSO regression.

Source: Author's elaboration.

Based on these steps, this book provides a robust methodology to process and organise information from job portals. As a result, the Colombian vacancy database created this way has the following structure as detailed in Table 6.2.[85]

[85] The following chapters provide a detailed descriptive analysis of the variables listed in Table 6.2.

Table 6.2. **Basic data structure**

Variable	Definition	Percentage of missing values
Job title	Short description about the job title	No missing values (mandatory field in the job advertisement)
Vacancy description	Detailed information about the profile required to fill the vacancy	No missing values (mandatory field in the job advertisement)
Labour experience	Dummy variable that takes the value of 1 if the vacancy (explicitly) requires any labour experience and 0 otherwise	No missing values (this variable takes the value of 0 if a vacancy does not say anything related to labour experience)
Number of vacancies	Number of job positions offered for each job advertisement	No missing values (mandatory field in the job advertisement)
Company name	Name of the company who published the job advertisement	Around 4.5% of job advertisements with missing values
Publication date	Starting date when the job advertisement was placed	Around 20.0% of job advertisements with missing values
Expiration date	Date when the job advertisement expires	Around 65.3% of job advertisements with missing values
Educational requirements	Set of dummy variables that identify the educational attainment required to fill the vacancy: a. primary; b. bachelor; c. lower vocational education; d. upper vocational education; e. undergraduate; f. specialisation; g. master's degree; h. doctoral degree. See Chapter 8.	Around 20.0% of job advertisements with missing values. After the imputation process, no observations had missing values in this variable.
Wage	Continuous variable that indicates the amount of money that the hired person will receive	Around 30.0% of job advertisements with missing values. After the imputation process, no observations had missing values in this variable.
Imputed wage	Continuous variable that indicates the amount of money (imputed) that the hired person will receive	No missing values
Type of contract	Set of dummy variables that identify the type of contract offered by the employer: a. fixed-term contract; b. indefinite duration contract; c. freelance; d. by activities	No missing values (this variable takes the value of 0 if a vacancy does not say anything related to type of contract)
Workday	Set of dummy variables that identify the workday offered by the employer: a. full-time; b. part-time; c. by hours	No missing values (this variable takes the value of 0 if a vacancy does not say anything related to workday)
City	Place where the vacancy is available	Around 1.2% of job advertisements with missing values
Sector ISIC	ISIC Code (2 digits if possible)	Around 39.1% of job advertisements with missing values

Variable	Definition	Percentage of missing values
Skills	Set of dummy variables that identify the skills required by employers according to ESCO	No missing values (this variable takes the value of 0 if a vacancy does not say anything related to skills)
Specific skills	Set of dummy variables that identify (country-specific) skills required by employers and are not listed in the ESCO dictionary	No missing values (this variable takes the value of 0 if a vacancy does not say anything related to specific skills)
ISCO Code	ISCO Code (4 digits if possible)	Around 4.2% of job advertisements with missing values

Source: Author's elaboration.

6.8. Conclusion

Job portals might be a rich source of detailed information concerning two of the most critical variables for human resources analysis, which are the skills and occupations required by employers. Nevertheless, to obtain consistent information for skills and occupational requirements from job advertisements, it is necessary the use of dictionaries or classifications, along with the implementation of more complex algorithms. Consequently, the first part of this chapter discussed and selected the best procedures to organise and categorise skills and occupational information.

First, for the Colombian case, information regarding skills is widespread in job advertisements. There is no national skills dictionary available to identify what words refer to in the job description for a certain skill; nevertheless, this chapter showed that the usage of international dictionaries such as the ESCO might facilitate building a methodology that identifies the skills demanded in each job advertisement for countries like Colombia. Moreover, with the help of text mining techniques, it is possible to determine country-specific skills that are not listed in the ESCO dictionary but are mentioned in the job vacancy description.

Second, job titles in vacancy advertisements can, potentially, be organised and coded into occupations. The categorisation of job titles into occupations is one of the most critical procedures because this variable summarises the main characteristics of labour demand (tasks and skills required) and it is a key input for other processes such as the imputation of wage and educational requirements. In this regard, the economic and statistic literature has developed

different methods and algorithms to classify job titles into occupations (manual coding, classifiers, machine learning algorithms, etc.). Each method has its advantages and disadvantages. Manual coding might ensure a relatively high level of accuracy (percentage of job titles coded correctly); however, given the large number of cases (job titles), manual classification is a time-consuming task. On the other hand, automatic coding might help to assign occupational codes over a relatively short period of time, but there might be a considerable number of observations misclassified. This accuracy rate depends on algorithm performance and database quality.

Among the automatic methods discussed in this chapter, there are two main statistical tools: machine learning algorithms and software classifiers (which contain a set of logic rules). The main disadvantage of machine learning algorithms is that they strongly depend on the training database (job titles previously coded). In Colombia, this kind of training database does not exist. Thus, software classifiers such as Cascot might be an excellent help in a context such as the Colombian one. However, Cascot does not successfully classify all the job titles.

Therefore, at least for the Colombian context, there does not exist a unique method that satisfactorily assigns occupational codes to job titles. Given the advantages and disadvantages of each approach, this document proposes a combination of techniques: 1) manual coding for the most common job titles; 2) a software classifier (Cascot) adapted to the Colombian context, and 3) an extension of a machine learning algorithm (nearest neighbour algorithm) that takes into account not only job titles, but also skill requirements. Additionally, a (short) manual revision of the automatic outputs is undertaken.

Once all relevant variables are cleaned and adequately categorised for job vacancy analysis, another critical issue is the duplication problem. As vacancy data are collected from different websites (some job advertisements can appear on more than one job board or even on the same job board), the second part of this chapter showed how to deal with duplicated records. Specifically, it was argued that an n-gram-based approach (which is not sensitive to minor changes in string variables), so far, is the best method to minimise this issue. However, it is essential to recognise that (with the techniques available today) there is no way (apart from using a time-consuming manual process) to demonstrate that all duplicated observations have been dropped.

Finally, relevant variables for the analysis of labour demand for skills, such as wages and educational requirements, contain missing values. These missing values can create biases in the study of labour demand. Thus, the third part of this chapter explained and used the hot-deck and LASSO methods to impute missing values into the "education required" and "wage" variables.

In summary, this chapter 1) provided a robust and detailed methodology to obtain, organise, and categorise skills and occupations from job portals for statistical analysis; and 2) showed how to deal with duplicated job advertisements and missing values for relevant variables. Thus, as an outcome of this and the previous chapter (Chapter 5), the vacancy database can now be tested.

7. Descriptive Analysis of the Vacancy Database

7.1. Introduction

From a theoretical point of view, Chapter 2 discussed what can be understood as skill mismatches and how this phenomenon might arise in a particular economy (e.g. due to imperfect information). Chapter 3 showed that this problem has specific relevance in countries like Colombia. Indeed, evidence from the labour market in this country suggests that skill mismatches might explain a substantial portion of high unemployment and informality rates. One factor that hinders the design of well-orientated public policies to tackle skill mismatches is the absence or scarcity of detailed labour market information. More specifically, given the high cost of collecting information about labour demand for skills through surveys, the composition and dynamics of Colombian labour demand are relatively unknown.

However, information regarding unmet labour demand can be collected from job portals with the implementation of relatively novel data mining techniques. These online sources might provide valuable information in real time and at a low cost for the analysis of labour demand and, thus, for the early identification of labour demand for skills as well as possible skill shortages. A better understanding of this information can provide proper information to training providers and policymakers and, in this way, might improve education and public policy designs to tackle issues of unemployment and informality.

This chapter describes the main characteristics of vacancy data collected and organised in Chapters 5 and 6. Section 7.2 shows the number of vacancies and job positions demanded on job portals. Then Section 7.3 displays the geographical coverage of the Colombian vacancy database. The fourth section provides a descriptive analysis of the labour demand for skills in Colombia, and analyses labour demand composition by education, occupation (at a four-digit level), new job titles, skills, and experience requirements. Section 7.5 shows labour demand organised by sectors. The sixth section analyses the

most notable trends in the Colombian labour demand by occupation: occupations with higher demand, occupations with a considerable increase, and occupations for which demand has decreased over time. Section 7.7 describes the distribution of wages offered by employers, and the last section describes other (secondary) characteristics of the vacancy database, such as the type of contract and the duration of vacancies.

7.2. Vacancy database composition

Chapters 5 and 6 have described the methods and challenges involved in obtaining and organising vacancy information from job portals. As a result of those methods, a Colombian vacancy database has been generated to be tested and analysed for public policy recommendations. The sample period runs from January 1, 2016 to December 31, 2018. Each observation in the database is a vacancy. By the definition that has been applied in this book, a vacancy can require one or more people (the total number of jobs or job placements available) (see Chapter 5). Following the above definition, the total number of observations (vacancies) in the database are 2,247,959, while the numbers of jobs are 5,720,513 (Table 7.1). Consequently, a vacancy advertisement on average contains 2.5 job placements.

As shown in Table 7.1, by volume, most of the vacancies (55.7%) and jobs advertised (total vacancies) come from Jobportal_a, followed by Jobportal_b (33.4%) and Jobportal_c (10.8%). Likewise, 65.2% of the total number of jobs originate from Jobportal_a, followed by Jobportal_b (23.7%) and Jobportal_c (10.9%)[86] (Section 7.4 will discuss the types of job titles posted on each job portal).

[86] This result reaffirms that websites such as Jobportal_c do not necessarily contain the majority of job advertisements, even when the website stated that it had 263,621 job vacancies on October 30, 2017 (see Chapter 5). As mentioned in Chapter 5, when clicking on some vacancy announcements on Jobportal_c, a new window redirected the search by opening another website where the vacancy was originally posted (e.g. Jobportal_b).

Table 7.1. **Total number of vacancies and job positions**

Source	Total vacancies		Total jobs	
	Number	Percentage	Number	Percentage
Jobportal_a	1,252,366	55.7%	3,734,835	65.2%
Jobportal_b	752,032	33.4%	1,358,911	23.7%
Jobportal_c	243,561	10.8%	626,767	10.9%
Total	2,247,959		5,720,513	

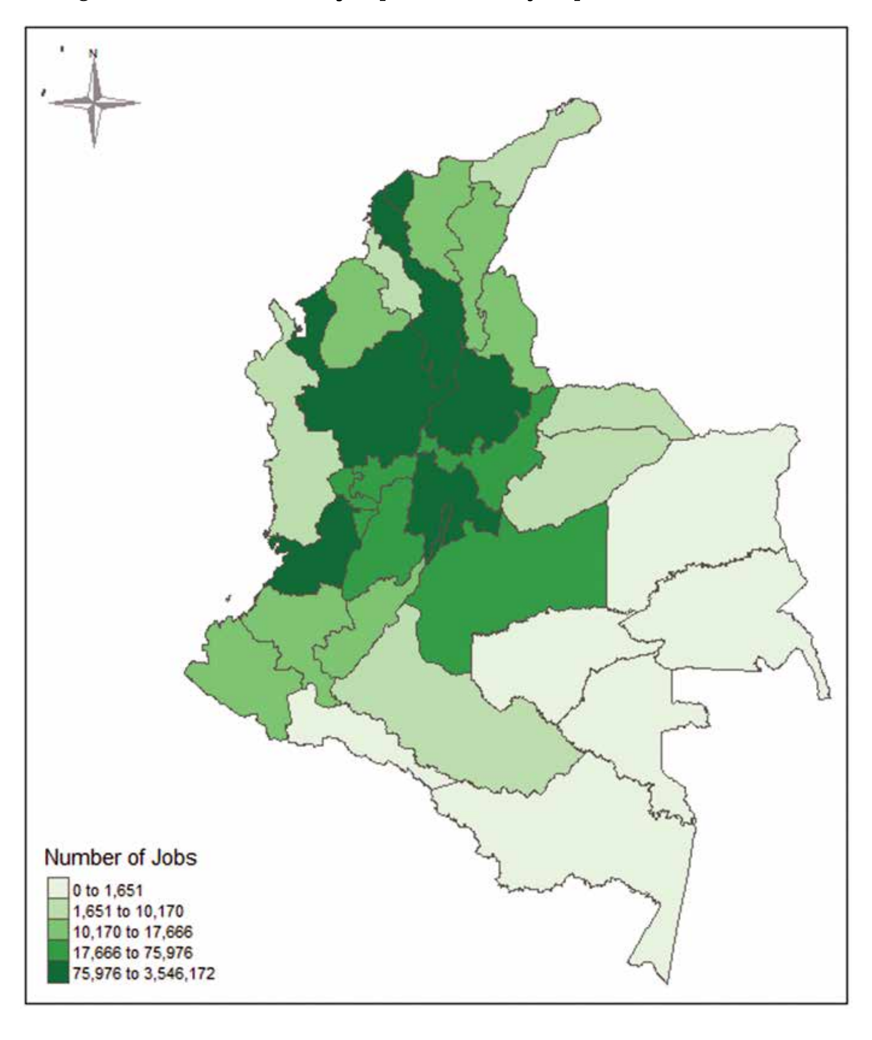
Source: Author's calculations based on vacancy information, 2016-2018.

7.3. Geographical distribution of vacancies and number of jobs

Figure 7.1 shows the distribution of vacancies in departments across the Colombian national territory from 2016 to 2018. Colombia is divided into 32 departments.[87] As can be observed, most vacancies and jobs are concentrated in the capital of the country (Bogotá). Indeed, 56.7% (1,276,410) of the total number of vacancies and 61.9% (3,546,172) of the total number of jobs were offered in Bogotá, while 7.9% of vacancies and 9.3% of jobs were available in Antioquia, and 15.2% of vacancies and 7.9% of jobs were offered in Bolívar. In contrast, the departments with fewer job placements are Vichada (228 job placements), Guainía (274 job placements), and Vaupes (75 job placements).

[87] Amazonas, Antioquia, Arauca, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Caquetá, Casanare, Cauca, César, Chocó, Córdoba, Cundinamarca, Guainía, Guaviare, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, San Andrés and Providencia, Santander, Sucre, Tolima, Valle del Cauca, Vaupés, and Vichada.

Figure 7.1. **Distribution of job placements by departments, 2016-2018**



Source: Author's calculations based on vacancy information and GEIH, 2016-2018.

Note: The ranges were chosen according to quintile distribution of job placements in the vacancy database.

It is unsurprising that more than half of the job placements in Colombia are concentrated in Bogotá, and departments like Vichada possess significantly fewer job placements. First, regarding the population, Bogotá is the biggest city in Colombia. According to the most recent figures published by the DANE,[88]

88 See http://www.dane.gov.co/files/investigaciones/poblacion/proyepobla06_20/Municipal_area_1985-2020.xls.

Bogotá has 8,281,030 inhabitants. This population represented approximately 16.4% of the total population in Colombia and 21.3% of the country's urban population in 2019. Additionally, Bogotá has 4,609,000 individuals in the economically active population (EAP). This number of people represented 18.6% of the total Colombian EAP and 23.6% of the urban EAP in Colombia in 2017.[89] Nevertheless, the above estimations do not consider that Bogotá attracts workers from its smaller surrounding cities. For instance, it is well-known that people from towns such as Soacha or Chía commute to Bogotá. Thus, by considering the surrounding cities,[90] the population of metropolitan Bogotá rises to 9,732,848, which represents 19.3% of the (total) population of Colombia and 25.1% of the country's urban population.

Given the economic concentration in Bogotá, this city produces 24.8% of the Colombian gross domestic product (GDP) (Valencia et al. 2016). Thus, it is logical to expect that the number of available vacancies is higher in Bogotá than elsewhere in the country. Likewise, the second largest department in terms of population and economic activity is Antioquia, followed by Cundinamarca, Atlántico, Bolívar, Valle del Cauca, and Santander. Therefore, it is also expected that these departments have a higher number of vacancies when compared to other Colombian departments. In line with this assumption, the departments of Vichada, Guainía, and Vaupes contributed only 0.3% of Colombia's GDP in 2017 (DANE 2018a, p. 4) and contained 0.33% of the total Colombian population in 2019.[91] Hence, it is unsurprising that those departments have a lower rate of job placements.[92]

Figure 7.2 shows Colombia's job distribution divided by EAP in each department from 2016 to 2017. The map does not include information from 2018 due to household data (GEIH) not being available at the moment.[93] The

---

89 See https://www.dane.gov.co/files/investigaciones/boletines/ech/ech/anexo_empleo_dic_17.xlsx.

90 Soacha, Facatativá, Chía, Zipaquirá, Mosquera, Madrid, Funza, Cajicá, Sibaté, Tocancipá, Tabio, La Calera, Sopó, Cota, Tenjo, El Rosal, Gachancipá, and Bojacá.

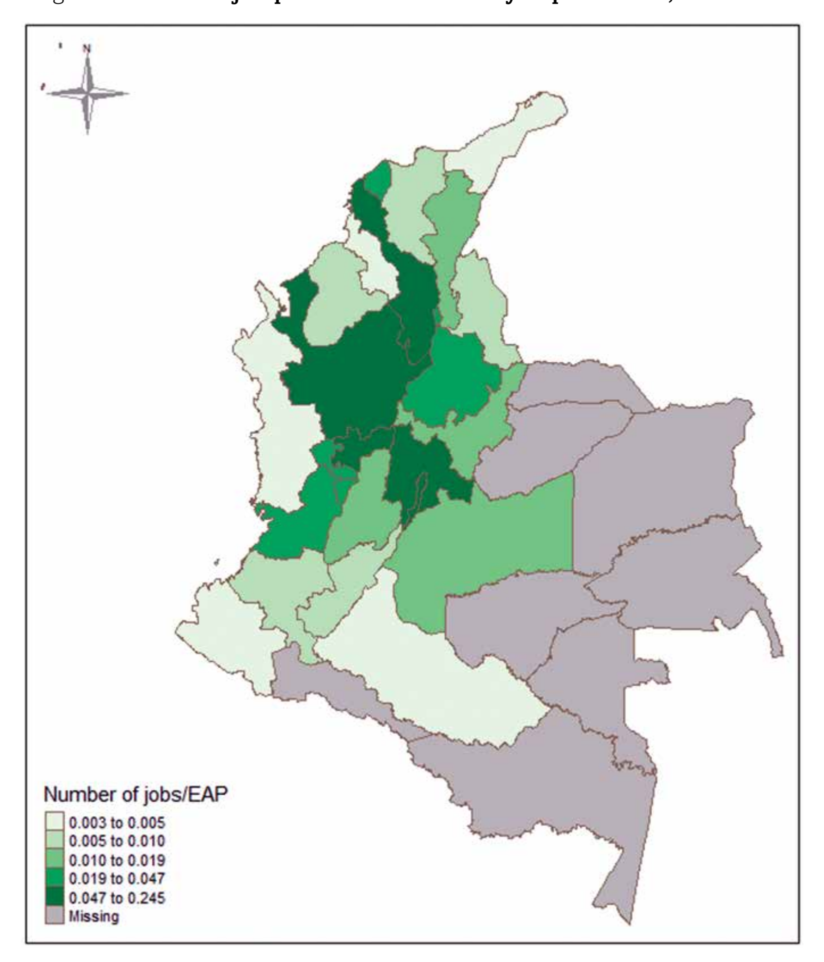
91 See http://www.dane.gov.co/files/investigaciones/poblacion/proyepobla06_20/Municipal_area_1985-2020.xls.

92 Chapter 8 will provide more detailed evidence about the external validity of the vacancy information.

93 This issue illustrates that there is a degree of delay between the release of household survey results and the problem researchers or policymakers want to analyse (Chapter 4).

first aspect to observe in the map is the presence of missing values; specifically, in the south-east zones of the Colombian territory. These missing values exist because there is no official information about the labour market (such as EAP and unemployed population, among others) in those departments.[94] Consequently, sources such as job portals might facilitate the provision of labour market information where it is difficult to carry out traditional methods (surveys).

Figure 7.2 **Ratio of job placements to EAP by departments, 2016-2017**



Source: Author's calculations based on vacancy information and GEIH, 2016-2017.

94 Due to problems of public order and the difficulty in accessing these areas of the country, the DANE collects information (since 2012) from the department's capital, but not from other cities in those south-easterly zones.

As shown in Figure 7.2, in Bogotá the ratio between job placements and EAP is 0.245, which means that for one job placement there are four employed or unemployed workers. For departments such as Antioquia, Cundinamarca, and Caldas and Valle del Cauca, the ratios are around 0.05 (for each job offer there are 20 workers), while for Bolívar the rate is 0.147 (for each job offer there are 6.7 workers).

Figures 7.1 and 7.2 show that the vacancy information is unevenly distributed across the national territory. Online job placements tend to be concentrated in specific zones such as Bogotá, Antioquia, Bolívar, etc. This concentration of data correlates with the relative economic importance of each department. Departments with a larger proportion of EAP and GDP also tend to have a relatively higher number of job placements. Thus, the geographical results of the vacancy information appear to reflect Colombia's economic and population dynamics.[95]

7.4. Labour demand for skills

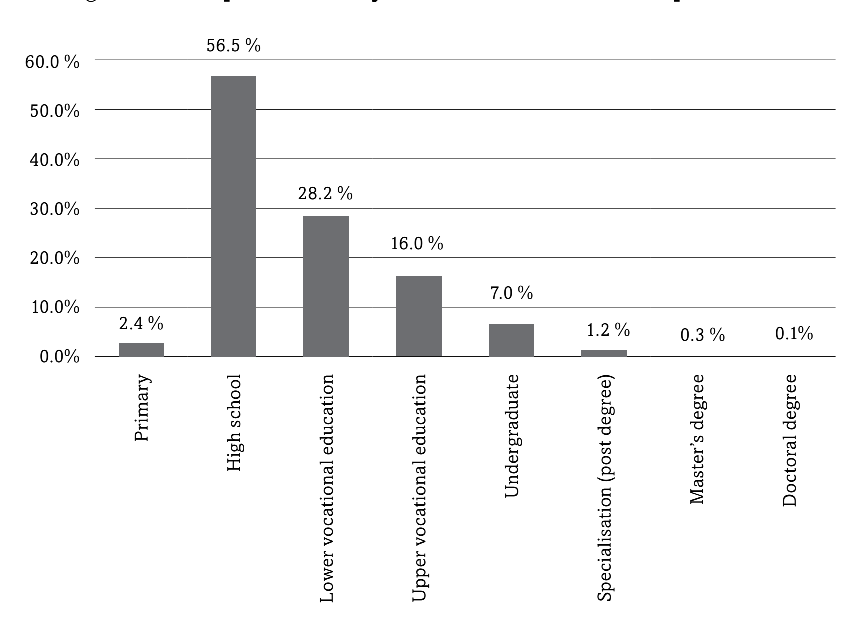
As discussed in Chapter 2, skill is a multi-dimensional concept. However, most skill definitions associate this concept with the task complexity attached to each job and the characteristics that each worker needs in order to successfully carry out tasks required in a certain job position. Reflecting on the definitions of skill from Chapter 2, skill is considered as any measurable quality that increases worker productivity and can be improved by training or development. Consequently, given the current sources of information available to analyse the labour market (job portals and household surveys), and the information provided by these sources, it is possible to analyse the Colombian labour demand by demanded education, skill, and experience (worker skills), and occupation (skills as job attributes) (see Chapter 2).

95 Potentially, the labour market analysis in this book can be disaggregated at the regional level. However, due to space limitations, (hereinafter) this document will present its results aggregated at the national level.

7.4.1. Educational requirements

Figure 7.3 shows the distribution of available jobs by educational requirements.[96] According to this figure, 56.5% of the job placements ask for a person with (at minimum) a high school degree, followed by lower (28.2%) and upper vocational educational requirements (16.0%). Despite using online sources (job portals), there are a considerable number of jobs that require people with a primary school and high school level of education, who tend to carry out low- or middle-skilled jobs. This evidence suggests (at least for the Colombian case) that companies do not only search for high-skilled workers when using job portals. As will be seen in more detail in this section, job placements posted on job portals cover a variety of low-, middle- and high-skilled jobs.

Figure 7.3. **Job placements by minimum educational requirements**



Source: Author's calculations based on vacancy information, 2016-2018.

96 As pointed out in Appendix B, employers might be indifferent about educational levels. For instance, a vacancy might require a person with a high school degree and lower vocational education level. In these cases, the educational dummy variables ("high school" and "lower_vocational _degree") take the value of one at the same time. For this reason, the sum of percentages in Figure 7.3 is more than 100%.

7.4.2. Occupational structure

With the occupational variable, it is possible to understand labour utilisation and the composition of an economy (high-, middle-, and low-skilled jobs); it also allows examining changes (such as job polarisation) in the labour force, and it serves as a guide for training providers and policymakers, among others.

Job portals provide job titles when a vacancy is posted online. As discussed in Chapter 6, there are different techniques available that might help to classify job titles from job portals into occupational groups. However, two concerns arise when using job title information from different job portals for the analysis of labour demand. First, job portals might be biased towards specific groups of occupations. Given that the vacancy database is composed of a group of main job portals in Colombia (see Chapter 5), the results might be biased due to one or more job portals only publishing vacancies for specific occupational groups. Second, the techniques carried out in Chapter 6 might misclassify some job titles, and thus the results regarding occupations might be affected.[97] This subsection provides evidence that the concerns mentioned above are not the case for the Colombian vacancy database.

Figure 7.4 shows a word cloud with the most common job titles for each job portal selected in Chapter 5. As can be observed, the most frequent job titles are “Call centre employees,” “Customer service” (“Cliente”), “Assistants” (“Auxiliar”), “Salespeople” (“Venta”), “Promoter,” etc. There are two aspects to highlight from Figure 7.4. First, the most demanded job titles correspond to low- or middle-skilled occupations. Second, the three job portals offer similar job positions. For instance, in all three job portals, one of the most common job titles is “Call centre employees.” This result suggests that the job portals selected in Chapter 5 are not biased to a specific market (i.e. high-skilled jobs such as managers or professionals).[98]

97 For instance, according to information from job portals and the techniques carried out in the previous chapter, one of the most demanded occupations might be “Actors.” It is not expected that an occupation such as “Actors” (or other occupations that usually do not have a big market) constitute an important share of the labour demand.

98 Chapter 8 will provide more evidence regarding the occupations demanded by job portals.

Figure 7.4. **Word cloud: Most frequent job titles by job portals**[99]

Jobportal_a



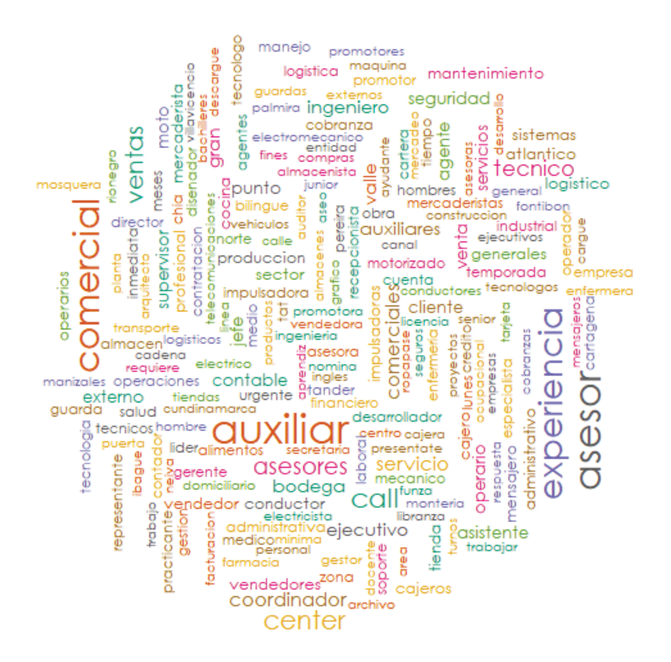
Jobportal_b



---

99 The text mining figures are presented in the Spanish language because it is the original language used on the Colombian job boards.



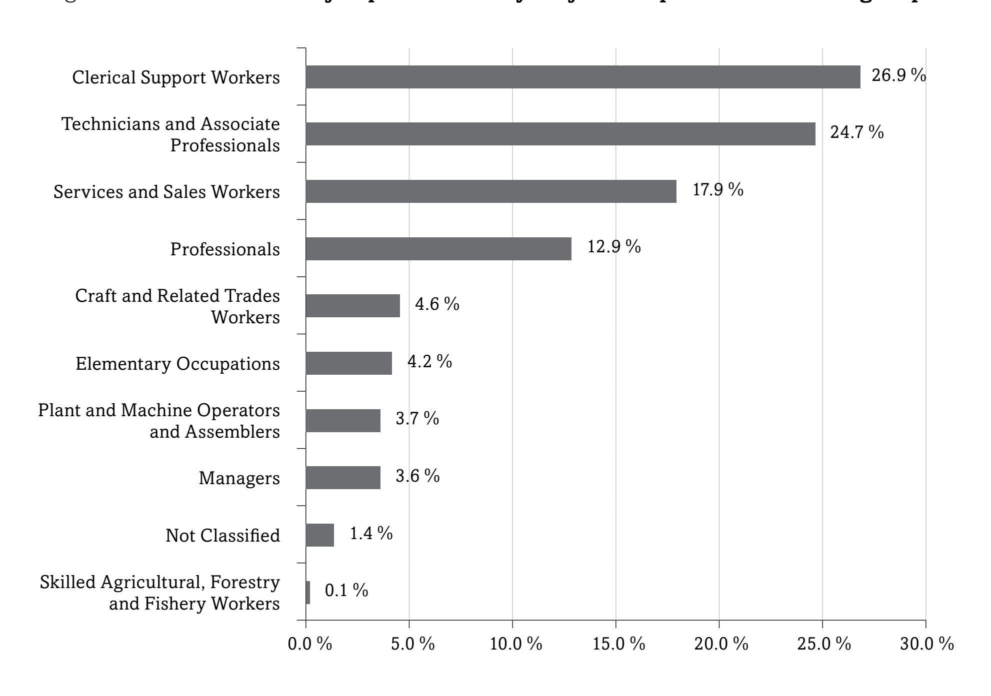


Source: Author's elaboration based on vacancy information, 2016-2018.

Figure 7.5 shows the distribution of job placements by their major occupational groups (aggregated at a 1-digit level ISCO-08). Around 26.9% of jobs demand "Clerical support workers," 24.7% demand "Technicians and associate professionals," and 17.9% required "Services and sales workers." Only 1.4% of job placements do not receive an occupational code ("Not classified"). These missing values correspond to observations that do not have enough or completely lack useful information in the job title (for instance, "students," "part-time job," etc.).

Figure 7.5 is useful because it shows the general structure of the vacancy database and, potentially, the Colombian labour demand structure. Moreover, these results reaffirm what was stated previously: companies do not only search for high-skilled workers when using job portals. However, as mentioned in Chapter 3, Big Data information unlock the opportunity to monitor job requirements at the disaggregated level (e.g. 4-digits occupation level).

Figure 7.5. **Distribution of job placements by major occupational ISCO-08 groups**



Source: Author's calculations based on vacancy information, 2016-2018.

Table 7.2 shows the occupational structure (at a four-digit level) of the labour demand. According to the vacancy data, the most required occupation in Colombia from 2016 to 2018 is "Commercial sales representatives" (15.4% of job placements), followed by "Telephone switchboard operators" (8.3% job placements), and "Stock clerks" (8.3% of job placements). These three occupations constitute around 32% of all job placements. Moreover, the Top 50 most demanded occupations form 78.2% of the Colombian labour demand. Consequently, according to the information from job portals, the most required occupations are related to sales, customer services, guards, and food preparation.

Another aspect to highlight is the presence of occupations related to technology and software development, such as "Information and communications technology user support technicians", "Information and communications technology operations technicians," and "Web and multimedia developers." This result confirms what was mentioned in Chapter 4: the labour demand for those occupations has dramatically increased during the last few years (Section 7.6 provides detailed evidence about labour demand trends).

Table 7.2. **Top 20 most demanded occupations in Colombia**

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
1	3322	Commercial sales representatives	878,503	15.4%
2	4223	Telephone switchboard operators	473,021	8.3%
3	4321	Stock clerks	472,076	8.3%
4	5223	Shop sales assistants	269,756	4.7%
5	5242	Sales demonstrators	235,481	4.1%
6	5230	Cashiers and ticket clerks	201,939	3.5%
7	4412	Mail carriers and sorting clerks	123,381	2.2%
8	5414	Security guards	111,717	2.0%
9	2411	Accountants	110,560	1.9%
10	1221	Sales and marketing managers	109,265	1.9%
11	4214	Debt-collectors and related workers	91,483	1.6%
12	9412	Kitchen helpers	75,535	1.3%
13	3343	Administrative and executive secretaries	73,364	1.3%
14	4110	General office clerks	69,875	1.2%
15	4322	Production clerks	67,997	1.2%
16	4311	Accounting and bookkeeping clerks	58,822	1.0%
17	8153	Sewing machine operators	54,628	1.0%
18	4222	Contact centre information clerks	50,337	0.9%
19	3312	Credit and loan officers	48,063	0.8%
20	5321	Health care assistants	45,279	0.8%

Source: Author's calculations based on vacancy information, 2016-2018.

Despite the potential theoretical bias of the information mentioned in Chapter 4, the results from Table 7.2 suggest (at least for the Colombian case) that job portals are not entirely focused on high-skilled occupations. Indeed, most of the categories listed in Table 7.2 are middle- (such as "Sales demonstrators") or low-skilled occupations ("Kitchen helpers"), which are the expected results for a developing economy such as Colombia. Additionally, the Top 20 most demanded occupations in Colombia do not present any unusual results. Occupations that usually do not have a big market (such as "Actors") do not constitute a large share of the Colombian labour demand. All the above results suggest that vacancy information from job portals might provide relevant information for a wide range of low-, middle-, and high-skilled occupations.

As mentioned above, with the job vacancies categorised into occupations, it is possible to identify the share of high-, middle-, and low-skilled occupations demanded in Colombia. For instance, the OECD (2017c) defines the following as high-skilled occupations (classified under the ISCO's major groups): 1) legislators, senior officials, and managers, 2) professionals, and 3) technicians and associate professionals. Similarly, middle-skilled jobs include: 4) clerks, 5) craft and related trade workers, and 6) plant and machine operators and assemblers, while low-skilled jobs include: 7) service workers and shop and market sales workers, 8) agricultural and fishery workers, and 9) elementary occupations.

Table 7.3 shows the distribution of jobs according to the above definitions: 22.5% (2,356,979) and 35.7% (2,011,352) of job placements correspond to low-skilled and middle-skilled occupations, respectively, while 41.8% of job placements are high-skilled occupations. It is important to notice that around 878,503 of job placements in the high-skilled group correspond to "Commercial sales representatives." Consequently, the high-skilled group is the most frequent due to the high demand for "Commercial sales representatives." Importantly, the results of Table 7.3 confirm that the vacancy information from job portals provide a high volume of information for low-, middle-, and high-skilled occupations.

Table 7.3. **Distribution of job placements by high-, middle-, and low-skilled occupations**

Classification	Number of jobs	Percentage
High skill	2,356,979	41.8%
Middle skill	2,011,352	35.7%
Low skill	1,269,604	22.5%

Source: Author's calculations based on vacancy information, 2016-2018.

7.4.3. New or specific job titles

As pointed out in Chapter 3, the labour market changes rapidly and new occupations (or job titles) emerge or disappear over time. This document defines as "new or specific job titles" those titles that are not in the ISCO list of occupational titles adapted for Colombia. Consequently, new or specific job titles

can correspond to new job titles or job titles that the ISCO list of occupational titles adapted for Colombia has not yet itemised.

As mentioned in Chapter 4, the early identification of these new labour demands has at least two economic benefits. On the one hand, it allows training providers to adapt their curricula and, therefore, it helps to adjust people's skills to suit labour market changes. On the other hand, the identification of emerging patterns in labour demand might provide occupational classifications with real-time information. Consequently, statistics and public policy designs based on an adapted occupational classification might provide more precise results according to different regional and sectorial contexts.[100]

Given that job portals generate detailed information on a daily basis, the systematic collection of data from these sources allows the identification of new job titles, providing thus key information to identify new or emerging occupations. By implementing text mining techniques (word clouds), it was possible to identify the most common words in those job titles without an occupational code (see Chapter 6). As mentioned in the previous chapter, a considerable percentage of those words do not describe a job position (e.g. "*Bachilleres,*" "undergraduate" in English). However, a deep visual inspection of these words reveals patterns (or phrases) that describe a job position. Given that this manual inspection to identify all the new or specific job titles for all the vacancy database is a time-consuming task, Table 7.4 presents the most recurrent new job titles identified in Colombia.[101]

---

[100] Provided the relevance of this topic for policy and education, institutions such as the O*NET have developed a methodology to identify, evaluate, and incorporate new and emerging occupations that have not yet been properly covered in the O*NET-SOC classification system (Dierdorff et al. 2009).

[101] It is important to note that these new or specific job titles might or might not be defined as new or specific occupations. As will be discussed in more detail in Chapter 10, further evaluation is required to determine whether a certain new job title corresponds to a new occupation (for instance, it is necessary to evaluate whether the new job title involves considerably different work than that performed by other job positions).

Table 7.4. **New job titles**

Job titles	Number of jobs
TAT vendors	52,849
Picking and packing assistants	8,652
CNC operators	2,840
Supervisor or specialist HSEQ	2,349
Baristas	1,715
Community manager	1,550
NIIF assistants, manager, or coordinator	1,532
Customer service social networks (Facebook, Twitter, etc.)	368
Cloud infrastructure engineer	169
SEO specialists	167
"*Maquetador web*" (web layout designer)	142
Datacentre operator	125
SSTA inspector	49
SQA professional	36
Influencer	23
Big Data specialist	14
Professor in Big Data	12
Bobcat operators	11

Source: Author's calculations based on vacancy information, 2016-2018.

It is worth noting the number of new job titles related to social networks and data management, such as "Cloud infrastructure engineer," "Professional SQA" (Software Quality Assurance/Advisor), "Influencer" (which is an industry expert who can influence other's behaviour through social networks, such as Twitter, Facebook, etc.), "Customer service social networks," "Big data specialist," and "Professor in Big Data", among others. However, it is not only in the IT sector that new job titles have emerged. Other job titles related to different activities have appeared too: "Supervisor or specialist HSEQ" (Health, Safety, Environment & Quality), "Baristas" (a person specialised in high-quality coffee, who creates new and different drinks based on their knowledge), and "TAT vendors" (store-to-store vendors, people who are considered as brand managers, and promote and sell products to local mini-markets).

Interesting occupational titles involve "CNC operators" or "Bobcat operators." In the ISCO-08 occupational titles provided by the DANE, these job titles are not listed, neither in Spanish nor in English; however, these job titles are listed in the UK version of ISCO-08. This result shows that some countries might faster identify emerging occupations compared to other countries, or that the arrival of some technologies occurs with certain delays for some developing countries like Colombia.

In general, new job titles involve new tasks or the use of new technologies. For instance, CNC operators programme and operate manufacturing machines. One difference with other operators is that CNC operators need to programme CNC machines to produce elaborate pieces of work. In contrast, certain kinds of jobs might be of particular interest for Colombia. For instance, this country is well-known as a producer of high standard coffee, and a considerable share of the Colombian economy depends on the performance of this product. Consequently, "Baristas" jobs might be essential job opportunities for Colombian workers, especially for informal and unemployed people. Baristas differ from other barman and similar occupations because the job of a barista requires a profound knowledge of high-quality coffee.

Thus, job portals are a rich source of changing information that requires the constant updating and adjusting of occupational classifications according to changes in the domestic labour market. Maintaining an updated occupational classification requires the continuous monitoring of occupations and new job titles, and might improve labour market matching and, hence, tackle informality and unemployment rates (see, for instance, Chapter 9).

7.4.4. The most in-demand skills (ESCO classifications)

As mentioned in the previous chapters, one of the most important characteristics of the vacancy database is that it might provide real-time and low-cost information about the most demanded skills in a particular economy. With the help of text mining techniques, it is possible to identify the skills explicitly demanded by employers according to the 13,485 skills listed in the ESCO, which is based on the principles of the European Qualifications Framework for lifelong learning (European Commission 2017) (see Section 6.2). Thanks to this well-known classification, it was not necessary to spend a considerable

amount of time in order to identify the words that describe skills in the vacancy database. Thus, this book avoids the use of a poorly defined list of skills and it uses an international categorisation instead.

Moreover, this dictionary groups the 13,485 skills into three groups: 1) knowledge,[102] 2) skill,[103] and 3) competence[104] (European Commission 2017). This categorisation provides a framework to analyse the patterns of skills demanded in Colombia. Additionally, the ESCO uses the concept of skill reusability (i.e. how widely a skill can be applied in different sectors or occupations) to divide the list of skills into four groups: 1) transversal,[105] 2) cross-sector,[106] 3) sector-specific,[107] and 4) occupation-specific[108] knowledge, skills, and competences. This definition is particularly useful because it allows identifying whether the Colombian labour demand requires general (transversal) skills or specific skills for an occupation or sector.

Following the above definitions, of the 13,485 skills listed in the ESCO, 4,051 were found in the vacancy database. In around 84.6% of the job advertisements mentioned, at least one word was related to skill information. For illustrative purposes, Table 7.5 shows the Top 20 skills most in demand in Colombia.

---

102 Knowledge refers to “the body of facts, principles, theories and practices that is related to a field of work or study. Knowledge is described as theoretical and/or factual and is the outcome of the assimilation of information through learning” (European Commission 2017, p. 6).

103 Skill is defined as “the ability to apply knowledge and use know-how to complete tasks and solve problems. Skills are described as cognitive (involving the use of logical, intuitive and creative thinking) or practical (involving manual dexterity and the use of methods, materials, tools and instruments)” (European Commission 2017, p. 6).

104 Competence is “the proven ability to use knowledge, skills and personal, social and/or methodological abilities in work or study situations, and in professional and personal development” (European Commission 2017, p. 6).

105 This category includes knowledge, skills, and competences that are important to a broad range of occupations and sectors. Usually, researchers refer to them as “core skills,” “basic skills,” etc. (see Chapter 2).

106 This category includes knowledge, skills, and competences that are necessary for different sectors. For instance, knowledge in “mechanics” is relevant for the automotive and textile industries (see European Commission 2017, p. 6).

107 This group refers to skills that are relevant for one sector but are required in different occupations. For instance, knowledge in “sales activities” is relevant for the marketing industry, but it is required for different occupations such as sales support assistant, shop manager, etc.

108 These skills tend to be used within one occupation or speciality. For instance, knowledge in “surgical instruments” is a relevant skill for surgical instrument makers.

Table 7.5. **Top 20 most demanded skills in Colombia**

Skills	Skill type	Skill reusability	Number of jobs	Percentage
Customer service	Knowledge	Sector-specific	827,705	14.50%
Communication	Knowledge	Cross-sector	480,653	8.40%
Work in teams	Skill/Competence	Transversal	322,457	5.60%
Work in shifts	Skill/Competence	Cross-sector	308,740	5.40%
Logistics	Knowledge	Cross-sector	208,013	3.60%
Blueprints	Knowledge	Cross-sector	169,579	3.00%
Telecommunication industry	Knowledge	Cross-sector	114,998	2.00%
Mechanics	Knowledge	Cross-sector	106,655	1.90%
English	Knowledge	Transversal	102,874	1.80%
Industrial engineering	Knowledge	Cross-sector	99,976	1.70%
Manage personnel	Knowledge	Cross-sector	96,579	1.70%
Customer insight	Knowledge	Sector-specific	94,318	1.60%
Electronics	Knowledge	Cross-sector	92,614	1.60%
Financial products	Knowledge	Cross-sector	66,990	1.20%
Accounting	Knowledge	Cross-sector	56,240	1.00%
Electricity	Knowledge	Cross-sector	42,391	0.70%
Telecommunications engineering	Knowledge	Cross-sector	38,967	0.70%
Sales activities	Knowledge	Sector-specific	37,411	0.70%
Sales strategies	Knowledge	Sector-specific	36,383	0.60%
Personal development	Knowledge	Cross-sector	35,160	0.60%

Source: Author's calculations based on vacancy information, 2016-2018.

As can be seen, the most in-demand skill is "Customer service" (14.5% of job advertisements), followed by "Communication" (8.4%), and "Work in teams" (5.6%). The second column of Table 7.5 shows that the most frequently demanded skill type in Colombia is knowledge. Additionally, according to the third column, most of the skills are cross-sector skills (14 out of 20 skills), followed by sector-specific and transversal skills (e.g. "Work in teams" and "English").

Importantly, the results of Table 7.5 are consistent with the occupational structure of Colombia (Table 7.2), where the most demanded occupations are "Commercial sales representatives," "Telephone switchboard operators," and "Stock clerks." Consequently, it is to be expected that the most frequently required skills are related to customer services, communication, and customer insight, among other skills.

Moreover, the ESCO groups the transversal skills into broader categories: values, ICT safety, application of knowledge, digital communication and collaboration, language, digital data processing, health and safety, problem-solving with digital tools, transversal skills/competences, attitudes and values, social interaction, thinking, digital competencies, numeracy and mathematics, working environment, and digital content creation. This aggregation provides an overview of the general structure of demanded transversal skills.[109] Table 7.6 contains the aggregated results of the demanded skills in Colombia. Social interaction skills (such as work in teams, manage personnel, assist customers, etc.) are the most demanded group, followed by language (mainly English) and thinking skills (develop working procedures, plan teamwork, perform market research, among others).

Table 7.6. **Skill groups demanded in Colombia**

Broader skill categories	Number of jobs	Percentage
Social interaction	895,530	15.7%
Language	109,708	1.9%
Thinking	46,865	0.8%
Numeracy and mathematics	25,340	0.4%
Health and safety	24,640	0.4%
Attitudes and values	23,881	0.4%
Problem-solving with digital tools	20,088	0.4%
Working environment	9,070	0.2%

Source: Author's calculations based on vacancy information, 2016-2018.

[109] For a detailed definition of each category see Larsen et al. (2018).

7.4.5. New or specific skills demanded in the Colombian labour market

As mentioned in Chapter 6, the ESCO is a useful dictionary to identify the skills required in the labour market in Europe. However, this dictionary might not fully identify the skills demanded in the Colombian labour market given that Colombian employers might demand different skills compared to Europe and because updating dictionaries to keep pace with changes in the labour market is challenging. Consequently, this book defines new or specific skills to address those skills that are not listed in the ESCO dictionary but are demanded in the Colombian labour market.

The identification of specific or new skills required in Colombia is relevant for tackling skill mismatch issues. With the implementation of new technologies, for instance, early identification of new skills required in the labour market might help people to adapt to Colombia-specific requirements and changes in the labour market, reducing thus unemployment and diverting people from joining the informal sector.

As described in Chapter 6, it is possible to identify potential words related to new and/or specific skills in the vacancy database. Given that it is necessary to carry out a careful visual inspection (which is a time-consuming task) to finally determine the words that describe skills (see Chapter 6), Table 7.7 shows twenty new or specific skills demanded in Colombia for illustration purposes.

Table 7.7. **Twenty new or specific skills demanded in Colombia**

Skills	Number of jobs
Packing or picking	67,493
SAP	21,378
Siigo	11,784
Pdv	7,360
Helisa	6,024
Scrum	4,219
Cosmetología	3,201
Apm	878
Perifoneos	858

Skills	Number of jobs
Mailings	536
Staad	336
Otdr	228
Rph	195
Kaizen	177
Fintech	176
Brandeo	149
Cloudera	138
Bigip	130
Rpgii	110
Ssst	98

Source: Author's calculations based on vacancy information, 2016-2018.

"Packing or picking" is an Anglicism that describes the process of gathering and placing individual components of an order into a box or envelope addressed to a recipient. These words are an example of terms that the Colombian labour market uses but are not incorporated into the ESCO dictionary. To identify these words, the vacancy database quantifies the real relevance of these logistic skills for the Colombian labour market. SAP (systems, applications, and products) is an integrated business management system designed to model and automate different areas of a company. According to the vacancy database, the use of this technology was necessary for 21,378 jobs between 2016 and 2018. Importantly, employers asked for knowledge in Siigo and Helisa (accounting and administrative softwares for enterprises). Clearly, these Colombian-specific softwares are not in the ESCO dictionary because the technology was developed and in demand in Colombia.

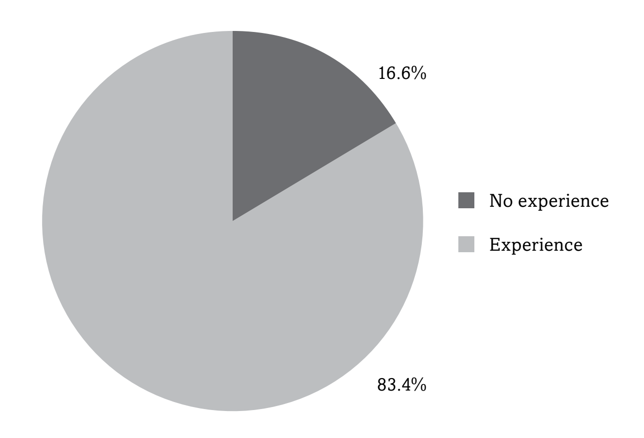
It is important to highlight here that some employer requirements—such as knowledge in Cloudera, Fintech, Mailings (email marketing)—have recently increased. In 2016, this knowledge was not demanded; however, by 2017 and 2018, these requirements began to appear in the Colombian labour market. This example shows that the analysis of information from job portals identifies changes in the labour demand for skills. Yet, it is also worth mentioning, that not only skill changes related to new technologies were found in the Colombian job market. For example, "*Cosmetología*" (cosmetology), "*Perifoneos*" (to

promote a product or business on the street with the help of a microphone) or "*Brandeo*" (an Anglicism of branding) are skills mentioned in the vacancy database that are not listed in the ESCO dictionary. Thus, job portals are a rich source of information to identify new or specific skills, which helps to update skills dictionaries such as the ESCO and improve education and training systems to meet specific requirements and changes in the domestic labour market.

7.4.6. Experience requirements

Regarding the experience requirements that employers are looking for in Colombia, 83.4% of the job placements explicitly require people with some work experience (see Figure 7.6). This result indicates that labour experience is an essential characteristic needed by workers when applying for most Colombian vacancies. While the median year for required experience is one year, it is important to note that this variable contains a large portion of missing values in the database. Indeed, 44.4% of the job placements that require some job experience do not report the specific years of work experience required.[110]

Figure 7.6. **Job placements by experience requirements**



Source: Author's calculations based on vacancy information, 2016-2018.

110 For instance, an employer might post a job advertisement in the following way: "a person with experience in photography is required to..." The variable "years of experience" can be imputed with the techniques explained in Chapter 5, and this imputation process will be part of a future research.

7.5. Demand by sector

An analysis of the vacancies by sector might serve to identify what skills or occupations are sector-specific or generic, which helps to address labour supply according to the needs of each industry. Thus, Table 7.8 shows the distribution of job placements by sector. As evidenced in the table, more than half of the job placements (around 55%) were coded according to ISIC revision 4 (division groups).

Table 7.8. **Job placements by sector**

ISCO rev4	Number of jobs	Percentage
Administrative and support service activities	2,070,156	36.2%
Wholesale and retail trade; repair of motor vehicles and motorcycles	338,387	5.9%
Professional, scientific, and technical activities	136,955	2.4%
Manufacturing	98,359	1.7%
Financial and insurance activities	97,351	1.7%
Construction	84,935	1.5%
Information and communication	74,502	1.3%
Transportation and storage	67,038	1.2%
Accommodation and food service activities	48,192	0.8%
Human health and social work activities	22,831	0.4%
Other service activities	14,661	0.3%
Arts, entertainment, and recreation	13,099	0.2%
Education	11,552	0.2%
Real estate activities	6,205	0.1%
Agriculture, forestry, and fishing	4,101	0.1%
Water supply; sewerage, waste management, and remediation activities	3,861	0.1%
Public administration and defence; compulsory social security	425	0.0%
Electricity, gas, steam, and air conditioning supply	308	0.0%
Activities of households as employers	6	0.0%
Not coded	2,627,589	45.9%
Total	**5,720,513**	

Source: Author's calculations based on vacancy information, 2016-2018.

As observed, companies related to “Administrative and support service activities” posted around 36.2% of the job positions, followed by “Wholesale and retail trade; repair of motor vehicles and motorcycles” (5.9%), and “Professional, scientific, and technical activities” (2.4%). The “Administrative and support service activities” category contains most of the job placements because this group includes companies related to “Temporary employment agency activities” and “Activities of call centres.” Temporary employment agencies act as a third party (intermediary) between companies and employees. They collect CVs and make their clients’ (employers) vacancies public. Consequently, if a vacancy is posted on job portals and the company’s name refers to a temporary employment agency, this does not mean that potential employees will work in the “Administrative and support service activities.” People who apply for those kinds of vacancies might end up working in other sectors (e.g. manufacturing). (Perhaps, information about the company’s name is in the description rather than in the company name variable. Thus, processing and identifying specific patterns in the job description might increase the number of observations of where people will work. However, this further development will be part of a future work).

On the other hand, it is expected that companies related to “Activities of call centres” have a considerable share of all job placements. As was shown in Table 7.2, there is a high demand for “Telephone switchboard operators,” among other related workers. Indeed, the results of Table 7.8 correlate with the results of Table 7.2 as the sectors with the highest number of job placements are related to the most demanded occupations. For instance, in Table 7.2, the most required occupations are related to “Sales,” “Customer services,” “Accountants,” and “Production clerks,” while the sectors with more job placements (apart from administrative and support service activities) are “Wholesale and retail trade,” “Manufacturing,” and “Financial and insurance activities.”

Another aspect to highlight is the relatively low frequency of job placements from sectors such as “Agriculture, forestry, and fishing,” “Public administration,” and “Defence companies,” etc. It was expected that these sectors would not have a high participation in the vacancy database because job portals (at least in Colombia) do not adequately cover rural zones where most of agriculture, forestry, and fishing companies operate, and, in addition, job portals are not a usual channel for posting vacancies related to public administration and defence, water supply, sewerage, and waste management, among other

activities. Instead, these vacancies are advertised on the website of individual companies; the scraping of that information will be part of a future work.

The issue of missing values in Table 7.8, as well as the high participation of "Temporary employment agency activities" might make it difficult to estimate the current level of labour demand by sector. Instead, information from job portals might be more useful for the identification of skills and possible skill shortages by industry. As can be seen in this table, there are a considerable number of observations for most sectors. This information might provide valuable insights regarding the most demanded generic and sector-specific skills and trends in the labour market by industry (see Chapter 8).

7.6. Trends in the labour demand

Although analysing the structure of labour demand is vital in order to know the kind of human resources required by employers, this analysis might not be sufficient to improve skills matching in the labour market if trends, seasonal changes, and business cycles are overlooked. The labour demand for certain occupations might increase over specific periods (i.e. quarters). For instance, in holiday periods, the need for "Hotel receptionist" might increase due to an increase in tourism. Moreover, the labour market is dynamic, and the labour demand for certain occupations or skills might increase/decrease over time. The analysis of labour demand cycles, seasons, and trends is of paramount importance because it enables training providers to adapt their curricula and train people in the required skills for technological changes, business cycles, etc.

Table 7.9 shows the distribution of vacancies and job positions across the period of analysis (2016-2018). As evidenced in the table, in 2016, the total number of vacancies and job positions was 688,477 and 1,746,762, respectively. In 2018, the total number of vacancies and job positions was 818,160 and 2,073,726, respectively. Consequently, the number of vacancies and the total number of jobs increased from 2016 to 2018 by about 15.8% and 15.7%, respectively. This increase in the number of job advertisements might correspond to economic growth in Colombia and the extended use of job portals to advertise job positions.[111]

[111] The next chapter provides more detailed evidence regarding this discussion.

Table 7.9. **Yearly distribution of vacancies and job positions**

Year	Total vacancies		Total jobs	
	Number	Percentage	Number	Percentage
2016	688,477	30.63%	1,746,762	30.54%
2017	741,322	32.98%	1,900,025	33.21%
2018	818,160	36.40%	2,073,726	36.25%
Total	2,247,959		5,720,513	

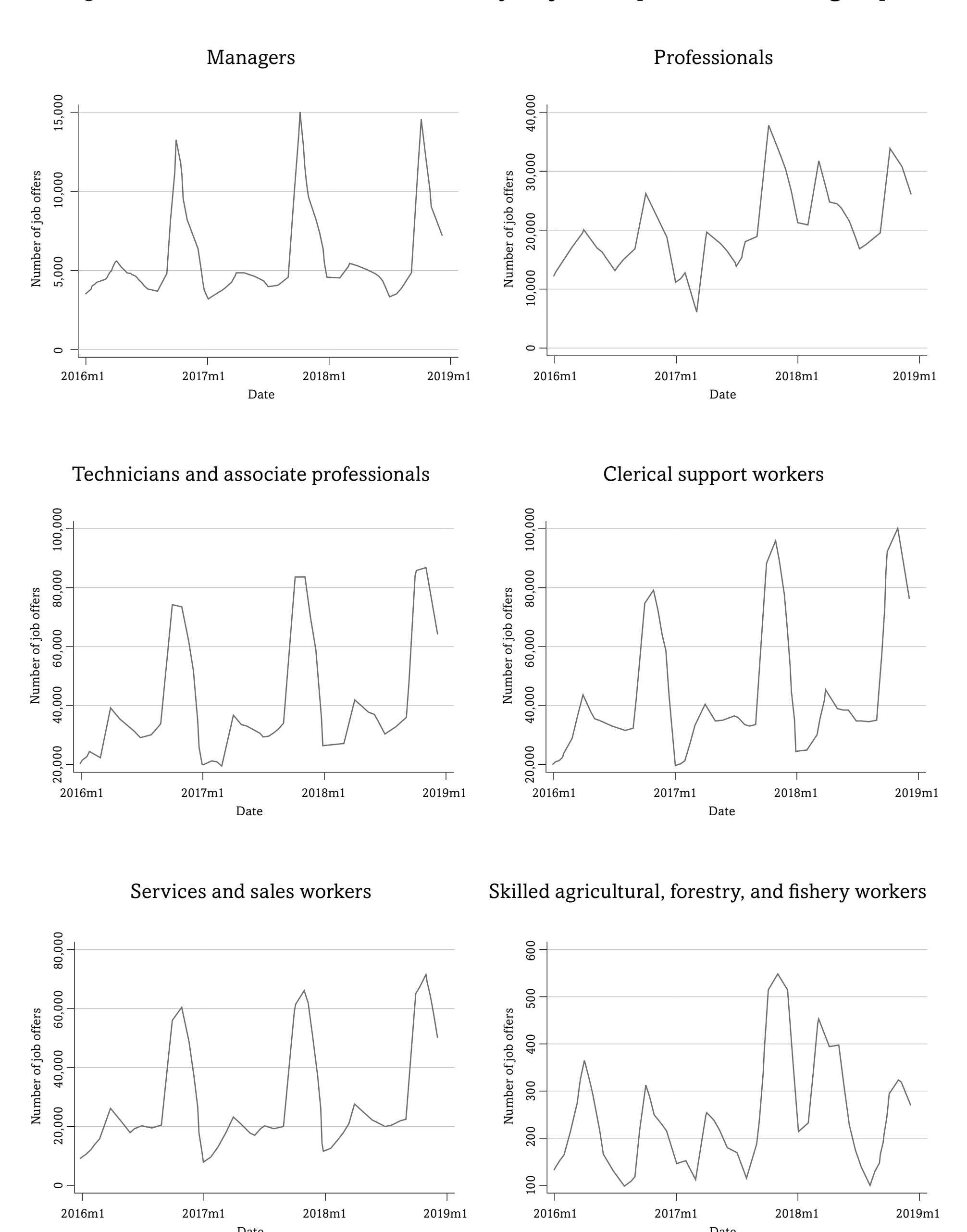
Source: Author's calculations based on vacancy information, 2016-2018.

Figure 7.7 shows the Colombian labour demand during the period of analysis for major occupational groups (one-digit level ISCO-08).[112] As seen in the figure, most of the major groups demonstrate an increase in labour demand between October and December, and a substantial decrease in demand between January and March.[113] In contrast, the labour demand for "Professionals" grew during the period of analysis. As mentioned in Subsection 7.4.2, aggregated results are useful because they provide an idea of global labour demand behaviour. However, analysing the results in a disaggregated way over time (for instance, at the four-digit level ISCO-08) produces summary measures of trends and amplifies labour demand behaviour.

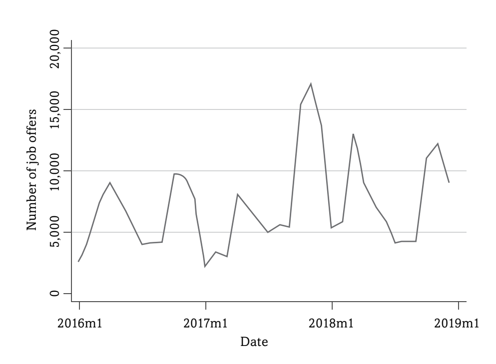
112 The labels "2016m1," "2017m1," etc., on the x-axis correspond to January 2016, January 2017, and so on.

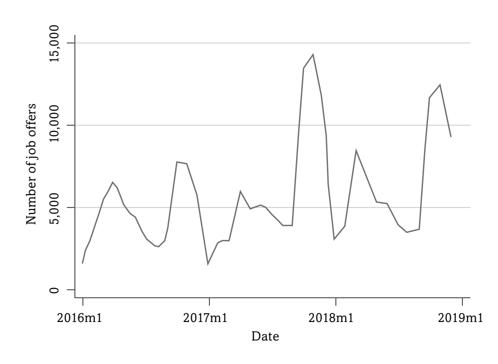
113 As will be seen in Chapter 8 in more detail, this cyclical behaviour correlates with the official unemployment statistics provided by the DANE, demonstrating that unemployment rates are relatively low between October and December, and higher between January and March. This result is due to the fact that companies hire people for the December season (when formal workers usually receive a Christmas bonus), and tourism, among other economic activities, considerably increases.

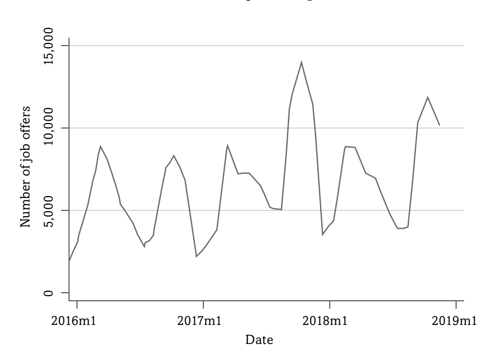
Figure 7.7. **Trends of the labour demand by major occupational ISCO-08 groups**











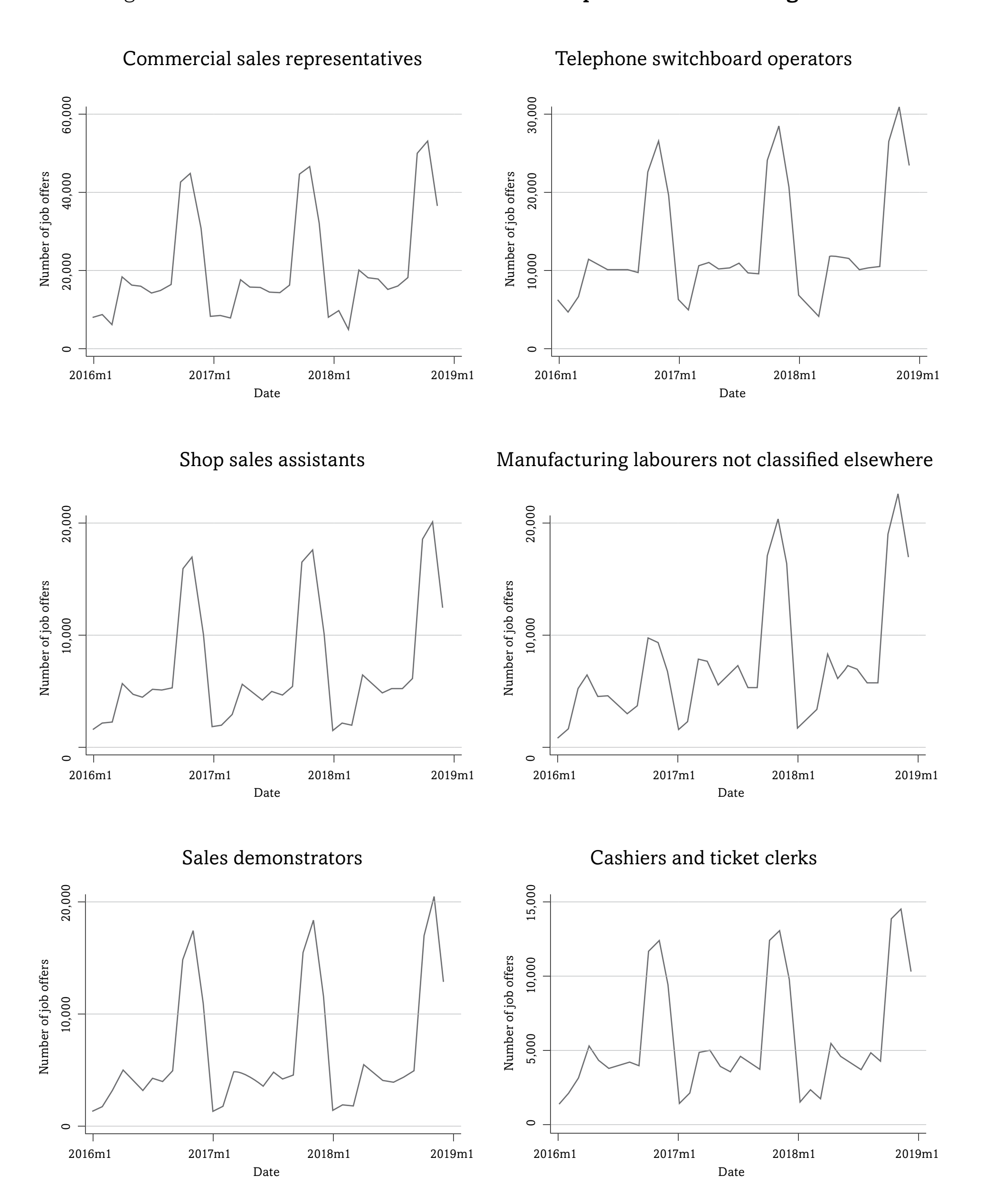
Source: Author's calculations based on vacancy information, 2016-2019.

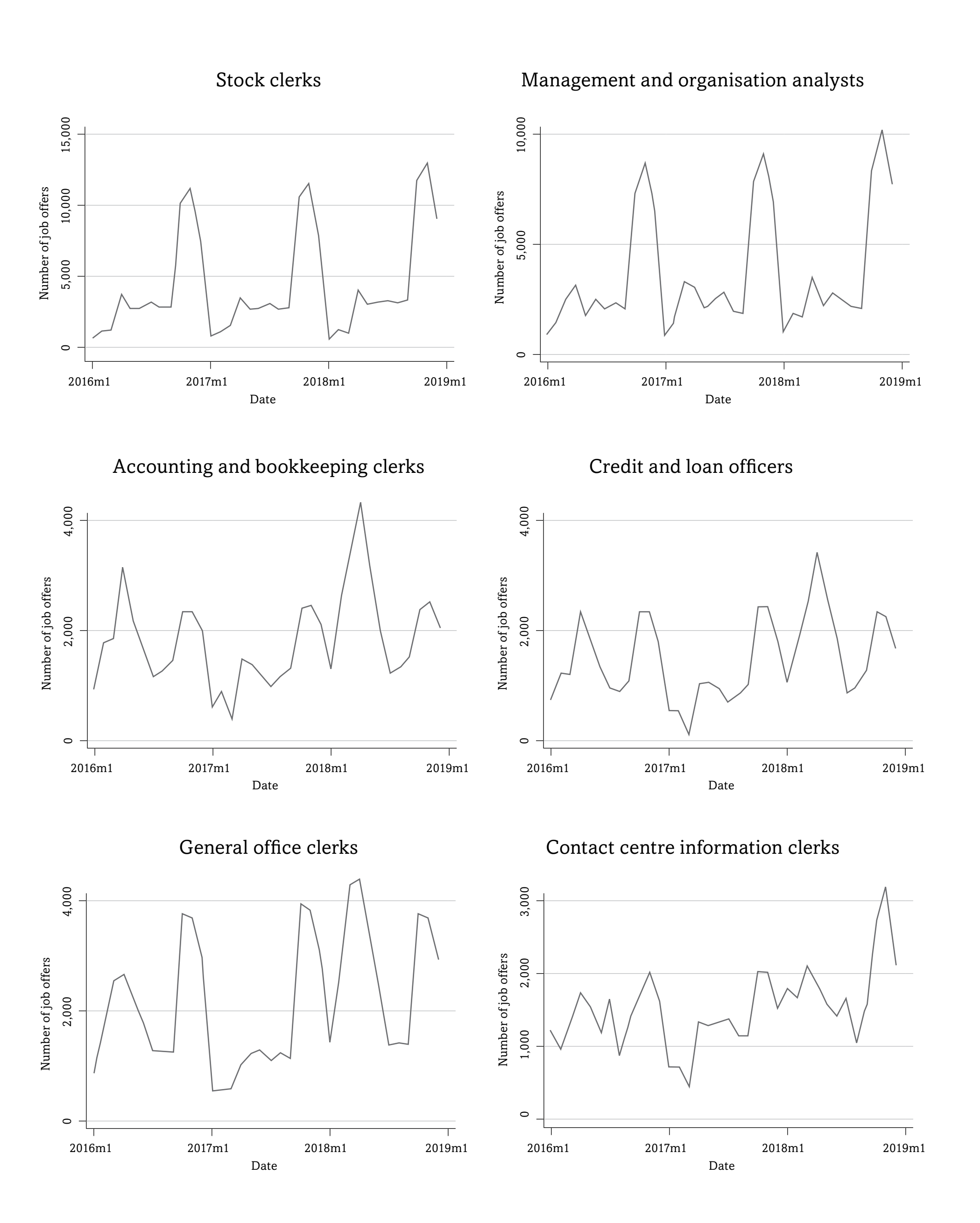
Figure 7.8, Figure 7.9, and Figure 7.10 show the most notable trends by occupational groups (the graphs for each occupational group [304] are available upon request). The charts are divided into three groups: Figure 7.8 shows the trends of occupations with a higher demand; Figure 7.9 presents occupations with a large increase in labour demand during the period of analysis, and Figure 7.10 displays occupations whose demand has decreased.

As can be observed in Figure 7.8, occupations with relatively higher demand tend to have a similar cyclical pattern over time: a remarkable increase in labour demand between October and December, and a sharp decrease between January and March. Alternatively, the labour demand for occupations in Figure 7.8 slightly increased during the period of analysis. In addition, there are also occupations that are always in high demand and, thus, do not exhibit a large increase in the last quarter of the year and a decrease in the first quarter. For instance, labour

demand for occupations such as "Accounting and bookkeeping clerks," "Credit and loan officers," "General office clerks," and "Contact centre information clerks" generally increase in the first and sometimes in the last quarter of the year.

Figure 7.8. **Trends of the most demanded occupations at a four-digit level**

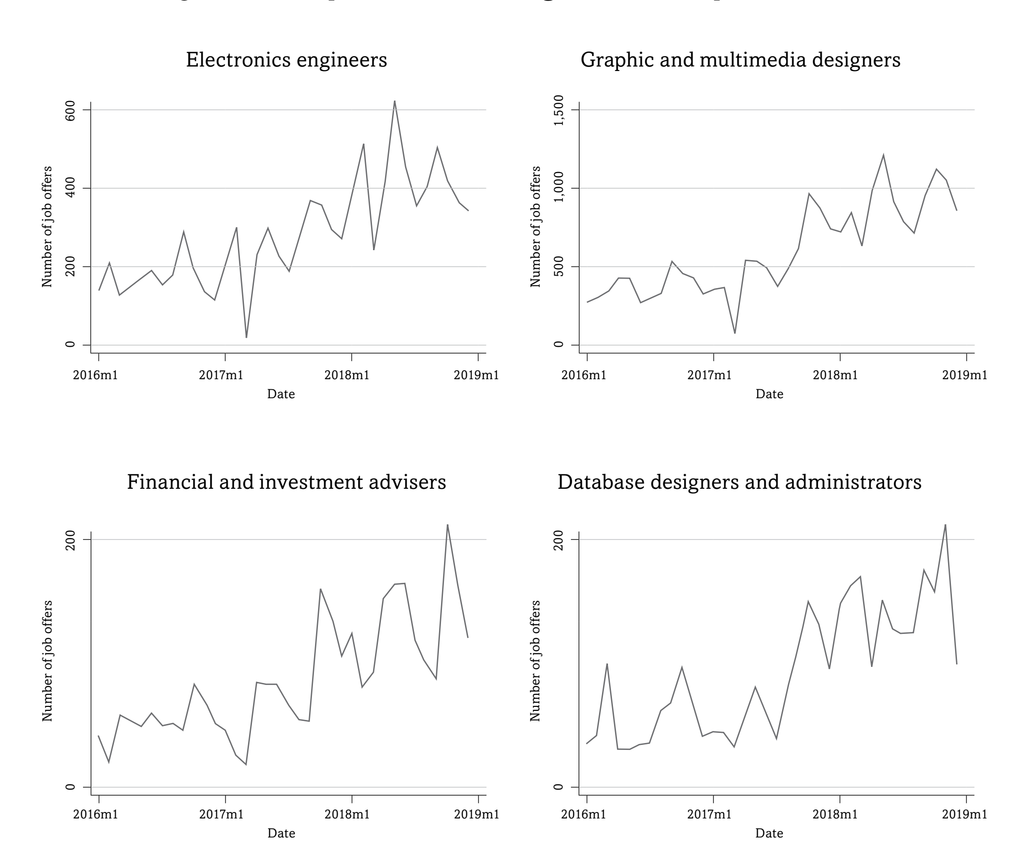


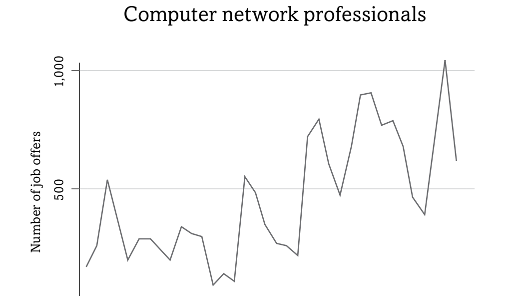


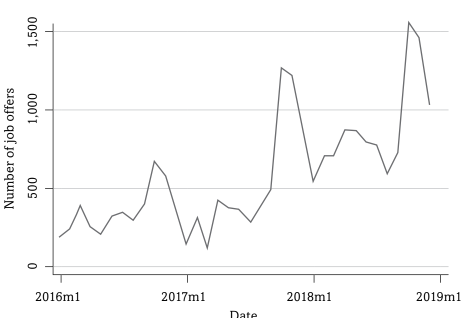
Source: Author's calculations based on vacancy information, 2016-2019.

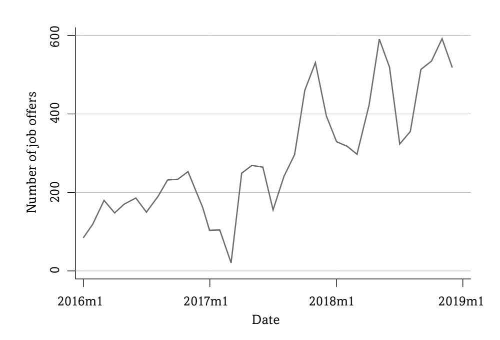
Figure 7.9 presents occupations with a considerable increase in labour demand. Despite the relatively short period of analysis (three years), it is possible to observe a growing trend of demand for "Electronics engineers," "Graphic and multimedia designers," "Financial and investment advisers," "Database designers and administrators," "Computer network professionals," "Electronics engineering technicians," "Real estate agents and property managers," and "Information and communications technology user support technicians," among others. These results suggest that in Colombia the labour demand for occupations related to technology, finance, and the real estate market is either rapidly growing or the companies that demand those occupations have increased their use of job portals.

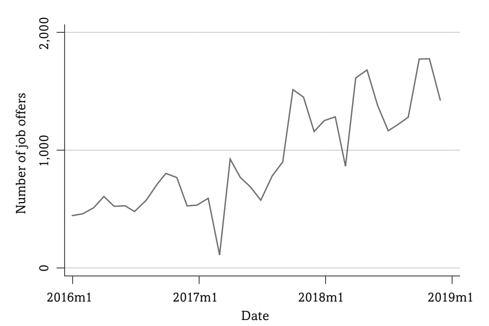
Figure 7.9. **Occupations at a four-digit level with a positive trend**











Source: Author's calculations based on vacancy information, 2016-2019.

Conversely, the labour demand for some occupations has decreased over time. Figure 7.10 presents occupations with a negative trend during the studied period (the results were grouped yearly in order to observe a clearer pattern). As can be seen in the figure, the demand for occupations such as "Cleaners and helpers in offices, hotels, and other establishments," "Waiters," "Receptionists (general)," "Dentists," among others, has decreased from 2016 to 2018. For instance, the labour demand for "Cleaners and helpers in offices, hotels and other establishments" decreased from 7,546 job placements in 2016 to 4,622 job placements in 2018. However, overall, there are relatively few occupations for which demand has decreased over time.

Figure 7.10. **Occupations at a four-digit level with a negative trend**

Cleaners and helpers in offices, hotels, and other establishments

Waiters

Receptionists (general)

Dentists

Source: Author's calculations based on vacancy information, 2016-2018.

The numbers in Figure 7.10 show that there is no dramatic decrease in labour demand for specific occupational groups. These results contrast with Figure 7.9 where the labour demand for certain occupational groups has dramatically increased. Two factors might explain the relatively high increase and the slight decrease of labour demand for particular groups of occupations during the period of analysis. First, as mentioned in Chapter 4, the use of job portals (and the internet, in general) has increased over the last decades. Consequently, as the number of job portal users increases, so will the number of vacancies posted on the internet. However, interestingly, the labour demand for certain occupations has decreased despite the increase in job portal usage. Thus, increased internet usage might soften the fall of job placements for particular

occupations, while this phenomenon intensifies the rise in job placements for other occupations.

Second, it has been widely reported that over the last decades there has been a skill-based technological change, which has increased labour demand and wages for skilled labour (Autor, Katz, and Krueger 1998). Thus, the remarkable increase in labour demand for occupations such as "Graphic and multimedia designers" and "Computer network professionals", among others (Figure 7.9), is a product of this technological change (i.e. structural change). Nevertheless, the "destruction" of labour demand for certain occupations is a process that might require a relatively long period. For instance, companies might adopt technologies that replace some human labour; however, making this transition requires a considerable interval. Companies need time to adapt their production process to new technologies, and legislation exists that protects job positions against (massive) layoffs or other abrupt changes. Thus, falls in the labour demand by occupation might not occur abruptly.

7.7. Wages

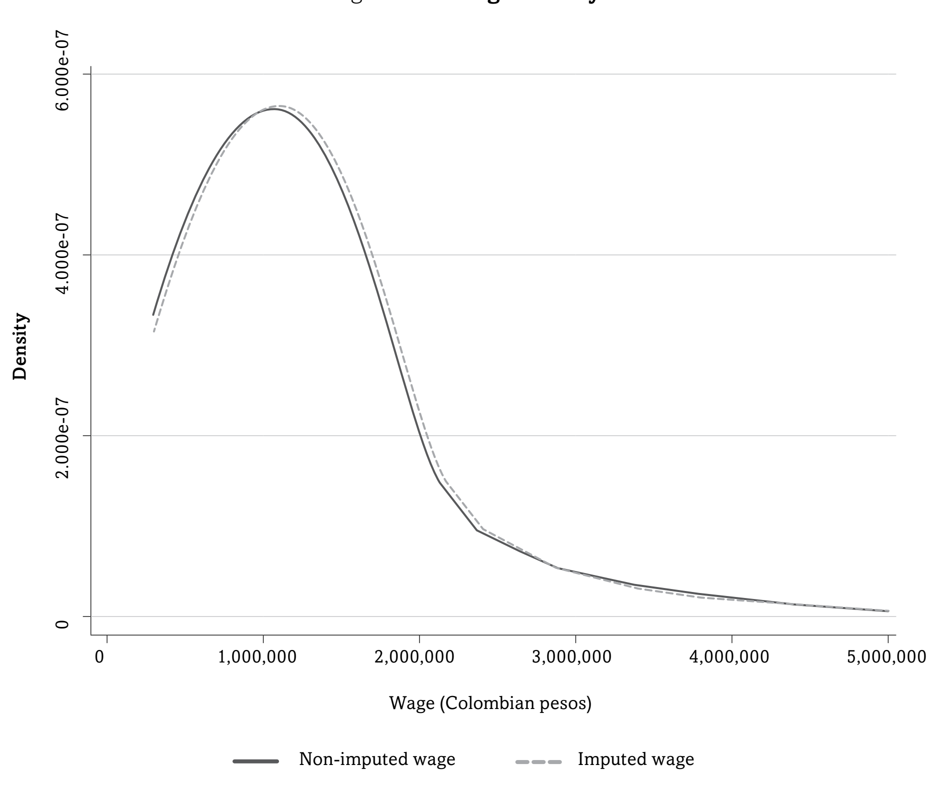
The analysis of the number of jobs posted in an economy is necessary but not enough to determine whether skill mismatches can be reduced. Jobs can be available; however, the wages of those jobs might not be high enough to create a labour supply to satisfy the labour demand. This variable helps to investigate whether the vacancies posted on job portals can offer wages that attract informal workers and the unemployed into formal jobs, and, at the same time, helps to determine possible skill mismatches (see Chapter 9).

Figure 7.11 shows the distribution of monthly wages from the vacancy database. The solid line represents the "wage" variable without any imputation process, while the dashed line represents the "imputed wage" variable (see Chapter 6). Both the imputed and non-imputed wage variables have a similar distribution because most jobs pay a salary between the legal monthly minimum wage[114] and 1,500,000 pesos (around £375). Indeed, the average

[114] In Colombia, the monthly minimum wage was 689,454 Colombian pesos (around £170) in 2016, $737,717 (around £184) in 2017, and $781,242 (around £195) in 2018.

figures for non-imputed and imputed wages are 1,059,667 pesos (around £265) and 1,102,200 pesos (around £275), respectively. These results reveal two facts. First, differences between non-imputed and imputed wages are minimal. Consequently, using imputed wages in the following chapters does not add considerable noise or bias to the statistical analysis; on the contrary, it enables an analysis of all vacancy observations. Second, the distribution of wages is consistent with the results from previous sections: a high proportion of jobs correspond to low- and middle-skilled occupations. Hence, the data are expected to have a right-skewed distribution (a high concentration of low wages) as in Figure 7.11.[115]

Figure 7.11. **Wage density**



Source: Author's calculations based on vacancy information, 2016-2018.

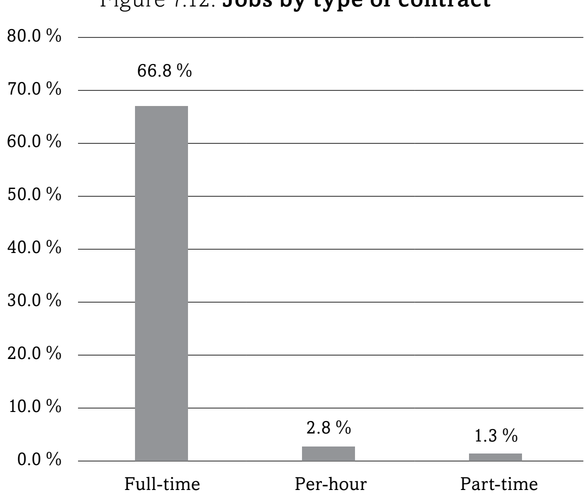
115 Chapter 8 provides more evidence about the consistency of the wage variable.

7.8. Other characteristics of the vacancy database

As mentioned in previous chapters, information from job portals is a rich source for identifying different characteristics of the labour demand. Some of those characteristics are not directly related to the labour demand for skills. However, this non-related skill information might provide more evidence regarding the consistency of the vacancy database, and it might be useful to tackle skill mismatches for a specific population or type of jobs. For illustration purposes, this section presents some of the most relevant characteristics of the vacancy database that are not directly related to skills information, such as type of contract offered and vacancy duration.

Figure 7.12 shows the distribution of jobs by type of contract. It is worth mentioning that some employers do not specify the kind of contract being offered. Moreover, and unlike the "education" variable, the variable "type of contract" is not imputed. Consequently, the sum of the percentage in Figure 7.12 is less than 100%: around 68.8% of the available jobs offer a full-time contract, while 2.8% and 1.3% of the available jobs offer a per-hour and part-time contract, respectively. This result suggests that information from job portals is not biased towards "irregular" jobs such as part-time work or per-hour jobs.

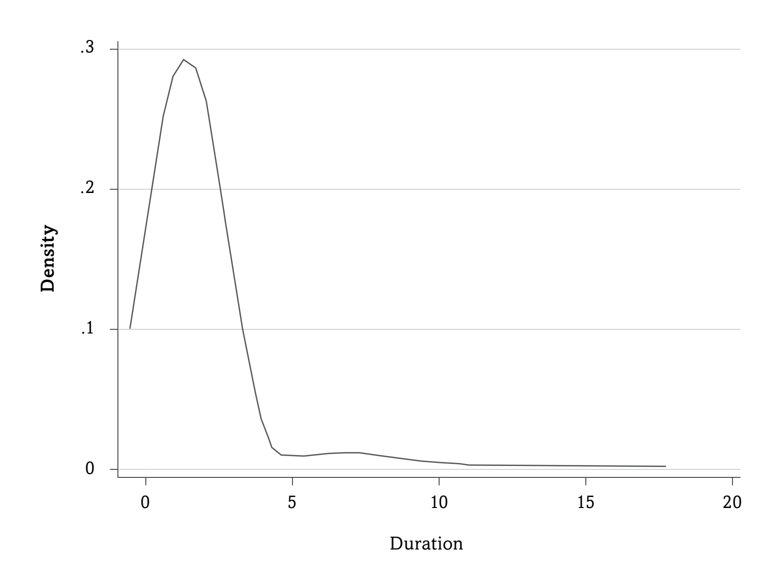
Figure 7.12. **Jobs by type of contract**



Source: Author's calculations based on vacancy information, 2016-2018.

Figure 7.13 shows the duration of job vacancy advertisements. This variable is the difference between the publication and the expiration date provided by employers in the job advertisement. The median advertising duration is 1.2 months. However, it is important to mention that 73% of the observations do not have information regarding their publication or expiration date. Despite the missing values, the results do not have atypical values. This result reaffirms that information provided by employers is consistent, and the problem of atypical or wrong values is minimum.

Figure 7.13. **Duration density (monthly)**



Source: Author's calculations based on vacancy information, 2016-2018.

7.9. Conclusion

The information from job portals has attracted the attention of researchers and policymakers, since Big Data seem to provide quick and relatively inexpensive access to information about employer requirements. Currently, for countries like Colombia, job portals are a unique source of labour demand information.

Much has been said about the advantages and limitations of using this information for labour demand analysis (see Chapter 4). For instance, given the online nature of these sources of information, job portal data might be biased towards high-skilled occupational groups. Nevertheless, most of the studies that have used job advertisements (printed or online) do not discuss the reliability of this information for labour demand analysis and public policy design (see Chapter 4). This chapter provides a descriptive analysis to start evaluating the results from the vacancy database, and its usefulness for tackling informality and unemployment problems in Colombia.

The sample period runs from January 1, 2016 to December 31, 2018. The main results of the analysis of the vacancy database show that 1) job vacancies are concentrated in Bogotá and the departments of Antioquia and Bolívar. These results are in agreement with other macroeconomic outputs. For instance, the capital (Bogotá) and its surrounding departments have the highest population and GDP rates; 2) most of the job positions require a person with at least a high school certificate; 3) in concordance with the previous result, most occupations in Colombia correspond to middle- ("Sales demonstrators") and low-skilled occupations ("Kitchen helpers"), which are the expected results for a developing economy such as Colombia. 4) This result also suggests that the job portals selected in Chapter 5 are not biased toward a specific market (e.g. high-skilled jobs, such as managers or professionals); thus, 5) job portals are a rich source of information for a continuous update of occupational classifications according to changes in the domestic labour market. For instance, among the most relevant new job titles found in the vacancy database are "TAT vendors," "CNC operators," and "Baristas."

In addition, regarding skill information in the vacancy database, the analysis shows that 6) the most demanded skills in the Colombian labour market include "Customer service" (knowledge), "Communication" (knowledge), and "Work in teams" (competence), which is consistent with demanded occupations; and 7) it is possible to identify new or specific skills such as Fintech, Mailings, and "*Perifoneos*," among others. Thus, it is possible to monitor changes and specific requirements in the domestic labour market at a low cost by using information from job portals, given that with a single (vacancy) database it is possible to analyse job attributes (of occupations in demand) and worker skill requirements.

Moreover, 8) the issue of missing values in the sector variable and the high participation of "Temporary employment agency activities" might make it difficult to estimate the current level of labour demand by sector. In these cases, information from job portals might be more useful for the identification of skills and possible skill shortages by sector.

Despite increased job portal usage, 9) it is possible to observe clear trends and seasons in labour demand; for instance, labour demand for certain occupations peaks in the last quarter of the year, while labour demand for occupations related to IT and other technologies has shown a steady growth. At the same time, labour demand for other occupations (such as "Cleaners and helpers in offices, hotels, and other establishments," "Waiters," and "Receptionists") decreased during the period of analysis.

The results regarding wages demonstrate two facts: 10) The differences between non-imputed and imputed wage distributions are minimal. Consequently, to use imputed wages in the following chapters does not add considerable noise or bias for statistical analysis; on the contrary, it allows analysing all of the vacancy observations. 11) The distribution of wages is consistent with occupations in demand. A high proportion of jobs correspond to low- and middle-skilled occupations and the distribution of wages is right-skewed (a high concentration of low wages).

The analysis of other characteristics of the vacancy database that are not directly related to labour demand for skills also shows two facts: 12) information provided by employers is consistent, meaning that issues such as outliers in the wage or vacancy duration variables are minimal; and 13) the vacancy database can provide different information such as what skill is most demanded by an occupation, or trends and seasonal changes in labour demand, which might serve as an input to tackle skill shortages for a certain population sample or certain types of jobs (see Chapter 9).

In general, the vacancy database provided detailed, real-time, and valuable information about the Colombian labour demand that previously was not possible to obtain from other sources (e.g. household surveys). Moreover, these initial results suggest that the vacancy database is consistent, or at least it does not contradict itself or external data, such as regional GDP, population, etc. However, a more detailed examination is necessary to draw conclusions about the reliability and representativeness of this vacancy database.



8. Internal and External Validity of the Vacancy Database

8.1. Introduction

The previous chapter described the main characteristics of the Colombian vacancy database from 2016 to 2018. However, these results do not provide enough evidence about the validity or reliability of vacancy data for addressing unemployment and informality problems in Colombia. As is the case with data collected with other methods (e.g. surveys), the data collected from online sources have some caveats that might affect the interpretation of results. Companies can post wrong or contradictory information; for instance, employers might request an engineering professional with just a high school diploma or a full-time engineering professional with an extremely low salary. Moreover, errors in posted information might arise from data mining processes. The algorithms created in the previous chapters might fail. For instance, the algorithm that looks for patterns in job descriptions might confuse some words, and incorrectly create variables of a university degree or job experience, among others. Consequently, errors or biases might arise in the information and affect the internal and external consistency of the vacancy database. Thus, this chapter tests the internal and external validity of the vacancy information.

Internal validity refers to the consistency of the variables within the vacancy database (Henson 2001; Streiner 2003). In ideal conditions, the results from a variable in the vacancy database should not contradict the findings from other variables in the same database; otherwise, the results will be unreliable. One straightforward way to address this issue is to compare the results of different but related variables. Therefore, the second section of this chapter tests the internal validity of the vacancy database via cross tabulations and wage distribution analysis.

Internal validity is a crucial aspect to consider before drawing any conclusions on labour demand from the vacancy database. Establishing result consistency from the vacancy database within a particular economic context

(external validity) is another relevant factor to consider before drawing any conclusions about Colombia's labour market (Kureková, Beblavy, and Thum, 2014). External validity, specifically, refers to possible biases or representativeness issues in the data (Rasmussen 2008; Stopher 2012).

Logically, all sources of information have limitations. For instance (as mentioned in Chapter 4), in Colombia, the current sectoral surveys carried out by the DANE do not provide detailed information about human capital, such as occupational structure or the skills required in each position. Web-based information might help to fill this gap. However, the online sources utilised for the database in this study also have limitations.

Given the nature of these online sources, job vacancy information might describe a particular segment of the labour market. The external validity of results depends on which kinds of vacancies are being published online for the country of interest. To test external validity, it is necessary to process and compare the results from other sources of information (e.g. household surveys) with the vacancy database results. Therefore, Section 3 discusses the representativeness of the Colombian vacancy database by categorising the household labour survey (GEIH) according to ISCO-08 categories, as well as comparing the Colombian vacancy data set with official national labour statistics.

Additionally, Sections 2 and 3 propose a framework to evaluate the representativeness of the vacancy database for each occupation at different levels of disaggregation (e.g. four-digit ISCO level):

- When testing the internal consistency of the information for a specific occupation, are there no errors or only minor errors?
- If yes, is the distribution of wages in the vacancy database for that particular occupation similar to the distribution of wages in the household survey?
- Can similar seasonal trends be observed in the level of employment in the household survey and in the level of job vacancies?
- Can opposite seasonal trends be seen in both the level of unemployment in the household survey and the level of job vacancies?
- Do lagged effects exist between the number of job advertisements and new hires?

This framework is particularly useful for countries like Colombia, where testing and comparing the representativeness of a vacancy database built from online sources is more challenging because labour demand information collected by traditional methods (such as vacancy surveys) is scarce.

8.2. Internal validity

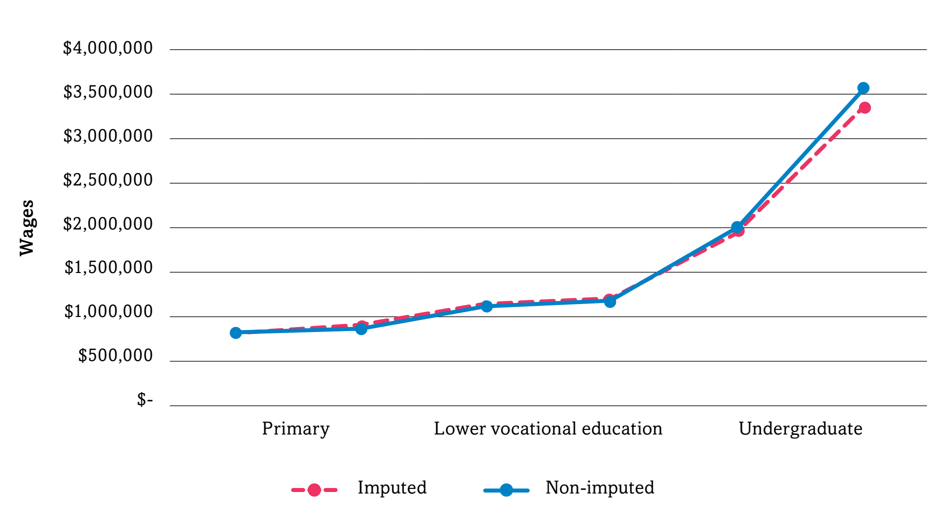
Establishing the internal validity or internal consistency of a database implies that the results from a variable should not contradict the findings from another variable (Henson 2001; Streiner 2003). If employers demand engineers or economists, for instance, most of the vacancies for those job positions should also require people with at least some university education level (when the educational level is mentioned in the job advertisement). Additionally, according to human capital theory, higher salaries should be positively correlated with a higher level of human capital (see Chapter 2); otherwise, the results would be contradictory. In this case, job portals might not be a reliable source of labour demand (skill mismatch) information, or the algorithms developed in Chapters 5 and 6 might be failing. Testing the internal validity of the vacancy database involves the comparison of different but correlated variables.

8.2.1. Wage distribution by groups

One straightforward way to prove the internal consistency of the vacancy database is comparing the average salary of different population groups. Usually, vacancies that require a person with a high level of education should pay higher wages than vacancies that ask for a person with a relatively low level of education (see Chapter 2). Figure 8.1 shows the average imputed and non-imputed salaries by educational level. As expected, vacancies that require people with a low level of education pay lower wages than vacancies that ask for people with a high level of education. On average, jobs that require a basic level of education (primary or high school) pay a salary of 829,000 Colombian pesos monthly (around £207), while jobs for undergraduates and postgraduates pay 1,975,040 pesos and 3,350,764 pesos (around £494 and £838), respectively. Moreover, as mentioned in the previous chapter, the differences between

imputed and non-imputed wages are minimal. This comparison suggests two facts: 1) The imputation process carried out in Chapter 6 does not considerably affect the wage distribution variable; hence, imputed wages (the whole database) can, potentially, be used for the analysis of labour demand; and 2) the vacancy information contains consistent results at least for the education and wage variables.

Figure 8.1. **Education and wages (Colombian pesos)**[116]



Source: Author's calculations based on vacancy information, 2016-2018.

Similar to the educational level variable, it is logical to expect that high-skilled jobs tend to pay higher salaries than low-skilled jobs. Figure 8.2 presents the average wages (imputed and non-imputed) that employers are willing to pay for high-, medium-, and low-skilled occupations. On average, the wage for a low-, medium-, and high-skilled occupation is around 970,000 (around £242), 1,034,000 (around £258), and 1,577,000 Colombian pesos (around £394), respectively. Moreover, the imputed and non-imputed wage variables overlap for each occupational group. Thus, there is a positive correlation between wages and the degree of complexity of an occupation.

[116] Given the relatively low frequency of specialisation, master's and doctoral degrees, these categories were grouped into a single category named "Postgraduate."

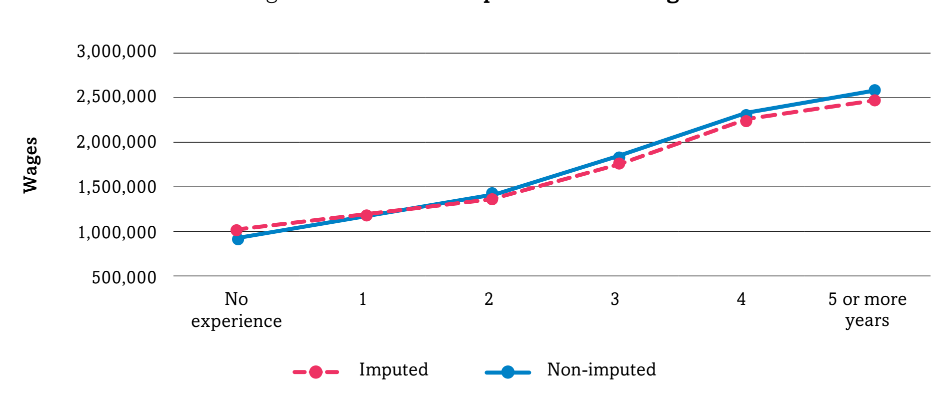
Figure 8.2. **Occupations and wages (Colombian pesos)**



Source: Author's calculations based on vacancy information, 2016-2018.

To provide more evidence regarding the consistency of human capital requirements and average wages in the vacancy database, Figure 8.3 presents the average wage (imputed and non-imputed) by experience requirements. Vacancies that do not require labour experience pay, on average, a salary of 1,045,000 Colombian pesos monthly (around £261), while vacancies that require five or more years of experience pay, on average, 2,457,000 pesos (around £838) monthly.

Figure 8.3. **Years of experience and wages**



Source: Author's calculations based on vacancy information, 2016-2018.

Therefore, the above evidence suggests that the information regarding human capital demand in Colombia is consistent with wage information. Consequently, this also means that the web scraping, text mining, imputation, and classification processes used to collect vacancy data provide consistent results with which to analyse the labour market. However, the wage variable only is an insufficient indicator to test internal data consistency given that vacancy distribution among groups is also required in order to determine the internal consistency of the vacancy database.

8.2.2. Vacancy distribution by groups

Similar to the wage variable, the distribution of vacancies should provide consistent results. As previously mentioned, if employers demand engineers, economists, or any other occupation that implicitly requires an undergraduate diploma, then most of the vacancies for such job positions should also demand people with at least some university education level. In addition, the skills listed in Chapter 7 should correspond to their related occupations; SQL programming skills, for instance, should correspond to programmers and other related occupations, being unlikely that programming skills would correspond to chefs, taxi drivers, or plumbers.

Table 8.1 reveals job distribution according to educational requirements for occupations following OECD categories (2017c). On the one hand, according to Column 1, around 41.1% of the jobs that require a basic education level (primary school) correspond to low-skilled occupations, while 56.2% and 2.7% correspond to medium-skilled and high-skilled occupations, respectively. On the other hand, only 1.6% of the jobs that require a postgraduate diploma correspond to low-skilled occupations, while 5.4% and 93.0% correspond to medium-skilled and high-skilled occupations, respectively. This result suggests that the information regarding human capital requirements in the vacancy database is consistent. Accordingly, in Table 8.1, the red zones indicate the lowest cell values; as the level of education increases, the percentage of low- and medium-skilled occupations decreases. The green zones indicate the highest cell values, and as the level of education increases, the percentage of high-skilled occupations increases.

Table 8.1. **Occupational structure by education**

Occupation	Primary	High school	Low vocational education	Higher vocational education	Undergraduate	Postgraduate
Low-skilled	41.1%	29.1%	15.0%	11.4%	6.2%	1.6%
Medium-skilled	56.2%	42.1%	32.8%	27.4%	14.0%	5.4%
High-skilled	2.7%	28.8%	52.2%	61.1%	79.8%	93.0%
Total	100%	100%	100%	100%	100%	100%

Source: Author's calculations based on vacancy information, 2016-2018.

Despite the sector variable containing a large number of missing observations (see Chapter 7), this variable might provide more evidence regarding the consistency of the vacancy database. A sector might demand different occupations that are not directly related to the main activity of the industry. For instance, the finance sector hires finance managers and finance analysts, among other associated occupations; however, this sector might also demand security guards and sales representatives. Notwithstanding the wide range of occupations required by each industry, differentiating patterns should exist between the occupational structure of one sector of the labour demand and another. The vacancy data should show, for instance, that the finance sector demands relatively more finance analysts than the agriculture sector; otherwise the vacancy information might contain considerable errors that can prevent a researcher from drawing academic or public policy recommendations.

Given the number of groups of occupations by industry, Table 8.2 shows some of the most notable cases of the labour demand occupational structure (at a four-digit level) by sector (at a one digit-level). For instance, Column 1 presents the ten most demanded occupations by companies related to "Real estate activities." As can be seen, the second most required occupation for this category is "Real estate agents and property managers," while in the other sectors this occupation is not frequently demanded. Companies related to "Accommodation and food service activities" frequently demand "Kitchen helpers," "Cleaners and helpers in offices, hotels, and other establishments," and "Stock clerks." It can be concluded from the table that occupations and sector variables have an expected correlation, which suggests that the occupational and industry variables, in general, provide consistent results.

Table 8.2. **Top 10 occupational labour skills in demand by sector**

#	Real estate activities	Accommodation and food service activities	Wholesale and retail trade; repair of motor vehicles and motorcycles	Manufacturing	Transportation and storage
1	Commercial sales representatives	Kitchen helpers	Commercial sales representatives	Commercial sales representatives	Stock clerks
2	Real estate agents and property managers	Cleaners and helpers in offices, hotels, and other establishments	Sales demonstrators	Sewing machine operators	Mail carriers and sorting clerks
3	Accountants	Stock clerks	Stock clerks	Cashiers and ticket clerks	Commercial sales representatives
4	Administrative and executive secretaries	Commercial sales representatives	Telephone switchboard operators	Stock clerks	Freight handlers
5	Telephone switchboard operators	Waiters	Security guards	Accountants	Building construction labourers
6	Building architects	Cooks	Cashiers and ticket clerks	Security guards	Security guards
7	Sales and marketing managers	General office clerks	Shop sales assistants	Shop sales assistants	Accountants
8	Stock clerks	Receptionists (general)	Waiters	Production clerks	Messengers, package deliverers, and luggage porters
9	Receptionists (general)	Cashiers and ticket clerks	Crane, hoist, and related plant operators	Services managers not classified elsewhere	Car, taxi, and van drivers
10	Survey and market research interviewers	Chefs	Accountants	Mail carriers and sorting clerks	Administrative and executive secretaries

Source: Author's calculations based on vacancy information, 2016-2018.

Skill information is one of the potential advantages of the vacancy database (see Chapters 6 and 7). Thus, it is essential to test the internal consistency of the skills variable. Testing this variable might be challenging because some skills (generic skills) are demanded regardless of the level of education, wage or occupation. Additionally, it might take a considerable time to test the

consistency of each skill.[117] To avoid these issues, ten skills explicitly related to an occupational group were chosen to test the internal validity of this variable. For instance, most of the jobs that require SQL or JavaScript programming skills should correspond to programmers and related occupations.

Table 8.3 shows the occupations with the highest demand for ten ESCO skills. For instance, the occupations with the highest demand for SQL programming skills are "Web and multimedia developers," followed by "Systems analysts and database designers and administrators." Similar occupations are demanded when employers require JavaScript skills. In contrast, when "Carpentry skills" are needed, the most frequently requested occupation is "Carpenters and joiners," followed by "Odd job persons," and "Mechanical engineering technicians." Additionally, "Generalist medical practitioners," "Nursing professionals," and "Specialist medical practitioners" are the most frequently demanded occupations when employers require epidemiology skills. This evidence suggests that skill information is consistent with the occupation variable, which, in turn, provides corresponding results with the educational level and wage variables.

Table 8.3. **Top 10 occupational skill categories**

#	SQL	JavaScript	Carpentry	Epidemiology	Mechanics
1	Web and multimedia developers	Web and multimedia developers	Carpenters and joiners	Generalist medical practitioners	Mechanical engineering technicians
2	Systems analysts	Systems analysts	Odd job persons	Nursing professionals	Electrical mechanics and fitters
3	Database designers and administrators	Engineering professionals not classified elsewhere	Mechanical engineering technicians	Specialist medical practitioners	Mining engineers, metallurgists, and related professionals
4	Information and communications technology user support technicians	Information and communications technology user support technicians	Stock clerks	Physiotherapists	Crane, hoist, and related plant operators

[117] Around 4,000 skills were identified in vacancy descriptions; see Chapter 7.

#	SQL	JavaScript	Carpentry	Epidemiology	Mechanics
5	Engineering professionals not classified elsewhere	Web technicians	Production clerks	Dentists	Mechanical engineers
6	Software developers	Software developers	Commercial sales representatives	Biologists, botanists, zoologists, and related professionals	Motor vehicle mechanics and repairers
7	Electronics engineers	Graphic and multimedia designers	Building construction labourers	Health professionals not classified elsewhere	Production clerks
8	Information and communications technology operations technicians	Electronics engineers	Sewing, embroidery, and related workers	Office supervisors	Mail carriers and sorting clerks
9	Web technicians	Building architects	Assemblers not classified elsewhere	Other artistic and cultural associate professionals	Heavy truck and lorry drivers
10	Electronics engineering technicians	Telecommunications engineering technicians	Information and communications technology installers and servicers	Chemists	Welders and flame cutters

Source: Author's calculations based on vacancy information, 2016-2018.

All the evidence presented above suggests that the vacancy database is internally consistent. However, it is important to note that every database, regardless of its sources, might have some errors.[118] For instance, Table 8.1 shows that around 2.7% (3,685 out of 136,479 observations) of the jobs that required education at primary school level correspond to high-skilled occupations. This result is suspicious because high-skilled jobs usually require a

[118] In household surveys, when people are asked about their wages, they can provide wrong information, or the interviewer might write an incorrect value. However, the depuration processes carried out by the statistics office guarantees that these measurement errors are minor and do not bias household survey results at a certain disaggregation level.

higher educational level. Indeed, a closer look at the vacancy database shows that a portion of these 3,685 jobs was misclassified.[119]

However, many mistakes are easy to identify and correct. In fact, one of the most critical advantages of scraping data directly from job portals is that researchers have the possibility to evaluate and correct possible mistakes in the gathered information. Algorithms might fail and might provide contradictory or inconsistent results; however, the quality of data created (i.e. dummy variables such as education and experience, among others) can be tested against the original data (i.e. job description, job title, etc.), and the algorithms can be easily refined until they provide a certain level of consistent results. For the Colombian vacancy database, the evidence shows that contradictory or inconsistent results are minor, and the magnitude of these measurement errors are not large enough to bias the educational, occupational, sectorial, skills, and wage analyses.

8.3. External validity

The previous section illustrates that the vacancy database provides consistent internal outcomes. Nevertheless, internal validity does not entirely prove the limits of the vacancy database. A database can provide consistent internal results, yet the data might not properly represent a population group (sample error); hence, academic or public policy conclusions drawn from that data might be biased.

Thus, the external validity or representativeness of a database is one of the essential elements to consider before drawing any conclusions based on that particular database (Stopher 2012; Rasmussen 2008). Despite the importance of testing the external validity of vacancy information, different authors have derived conclusions based on information from job portals without a careful analysis regarding data representativeness (see, for instance, Kennan et al. 2008; Backhaus 2004; and Kureková, Beblavy, and Thum, 2016). However,

[119] For instance, some of these jobs demanded "primary school teachers" and the text mining algorithm misunderstood it because the pattern "primary school" was in the job description, hence the educational requirement was wrongly assigned to "primary."

since this information does not come from a sampling frame, these sources may not be representative given the penetration of internet usage (Štefánik 2012). According to Carnevale, Jayasundera, and Repnikov (2014), the main source of bias in a job vacancy database might be due to differences in internet access among job applicants in terms of education level or skills.

Thus, more information does not guarantee better results. In consequence, not knowing the direction of the bias might provide the wrong conclusions or limit the scope of the studies. A possible bias in the collected information can affect vacancy analysis in two ways: 1) Job portals could publish only high-skilled jobs, while printed or voice-to-voice vacancies might correspond to middle- or low-skilled jobs. This possible source of bias might lead (in this case) to overestimating the labour demand for high-skilled jobs, and educational providers might saturate the labour market with more high-skilled people than the labour demand requires. In this document, this bias is named "selection bias." 2) The vacancies posted on job portals might not properly describe the characteristics (e.g. skills) required by employers. Jobs portals could tend to publish particular information to attract the attention of those who use the internet, while printed or voice-to-voice vacancies might publish different information (such as skills or educational requirements) to attract those people who use this medium to search for jobs. In this book, this bias is named "description bias."

Concerns regarding "description bias" were in part answered in the previous section. As observed, job requirements, such as skills, education, etc., correspond to the expected requirements for each occupation. "Description bias" seems implausible in the vacancy database because occupational requirements do not depend on the way the vacancy is advertised. For instance, the skills needed for a plumber do not change because the vacancy was posted online or transmitted voice-to-voice—the general tasks of a plumber are the same.[120]

However, vacancy data per se cannot answer when it is appropriate to provide more or less of a particular skill in response to labour demand. To accurately address this issue, it is necessary to identify any possible "selection

[120] Subsection 8.3.1.2 provides more evidence on this point, suggesting that "description bias" is not a predominant issue in the vacancy database. Thus, vacancy data can provide valuable answers about what people should be trained in at a low cost (time and money).

bias." Job portals might advertise more or fewer vacancies for a specific occupation regardless of the economic season or trends.[121]

Testing the "selection bias" might be challenging. As mentioned in Chapter 4, official labour demand surveys are characterised by a sampling frame (based on a census of people, companies, etc.), which ensures that the data and results are representative of a certain population. Consequently, given this statistical design, it is relatively easy to calculate the degree of representativeness in official household and sectorial surveys. Nevertheless, testing vacancy data representativeness is not an easy task given that this information is not collected based on a sampling frame. Ideally, in order to examine the data representativeness of information collected from job portals, an updated census of vacancies is required, which details the characteristics of human resource requirements. Nevertheless, carrying out this census is costly. Thus, countries like Colombia do not have a census of vacancies or any similar labour demand information to refer to. This absence of a vacancy census or survey makes it difficult to know the limits of information from job portals.

One way to address this issue is by comparing vacancy information with household surveys. Indeed, Štefánik (2012) compares the most popular job search website vacancies for tertiary education graduates in Slovakia with a labour force survey for the same educational group. As Štefánik (2012) points out, this approach assumes that occupational and sector structures in the vacancy database are similar to employment distribution by occupational and sector groups. According to this method, a vacancy database adequately represents labour demand information if there is a sufficiently high correlation with employment surveys. In aggregated terms, comparing vacancy data with household surveys can provide relevant insights regarding the representativeness of information from job portals. For instance, by comparing the number of vacancies with the level of employment over time, it is possible to determine whether job portal data adequately captures the behaviour of companies during economic cycles and seasons. It is expected, for instance, that the level of vacancies sharply increases at the end of each year given the increase in

[121] For instance, employers might opt to use job portals to collect CVs and store them in their databases (see Chapter 4) regardless of whether it is a period when more people are hired or not. Consequently, vacancy information from job portals might not be a useful source to determine trends, seasonal or cyclical changes in labour demand.

economic activities during that period, or the number of vacancies decreases during periods of economic recessions.

Moreover, the comparison between aggregated (one- or two-digit level) occupational structures of the vacancy database with occupational groups from household surveys might identify a possible under/overrepresentation of specific occupational groups in the vacancy database. At a one- or two-digit level of occupations (household surveys at a more disaggregated level such as a four-digit ISCO might have representativeness problems), both the vacancy and household data should have a similar occupational distribution if information from job portals adequately covers all occupational groups in the economy. Otherwise, vacancy information might over/underrepresent a particular occupational group.

One alternative explanation for the difference between the occupational structure of vacancy and household surveys might be that the labour market has a relatively high skill shortage problem. Given the existence of mismatches in the labour market, labour demand information might not coincide with labour supply information. This argument might justify why detailed comparisons between vacancy and household data are an improper method to test vacancy data representativeness. However, in aggregated terms (one or two occupational digit level) the differences between the labour structure of labour demand and supply might not be properly explained by the hypothesis of skill mismatches. For instance, a higher participation of "Professionals" (one-digit level ISCO, major group) in the vacancy database compared with information from household surveys would suggest, under the hypothesis of skill mismatches, that the country has a shortage of professionals of any kind. Nevertheless, this explanation does not seem plausible because if there were such evident skill shortages, the wages of professionals would be considerably higher, and the unemployment rate would be considerably lower than in other occupational groups. With such obvious evidence concerning labour market mismatch, education and training providers, the government, and, in general, people should react to this imbalance and correct the issue. For these reasons, the mismatch hypothesis might not explain occupational differences at a one- or two-digit level between vacancy and household survey information.

Thus, to compare the vacancy database at an aggregated level (i.e. major occupational groups) with the information from household surveys is the most

straightforward approach to identify possible biases in information from job portals. However, conducting a more detailed comparison to test the data representativeness of vacancy information between the vacancy database and a household survey might be problematic. In concordance with Kureková, Beblavy, and Thum (2014), household surveys provide information regarding labour supply, which is composed of the number of job matches (level of employment, see Chapter 2) and the number of unemployed people, while information from job portals is the total of the net and replacement labour demand.

Therefore, a direct comparison with household surveys at a detailed level (i.e. ISCO minor groups) might not be a suitable proxy to test data representativeness in the vacancy database. Besides, vacancy information might contain and reflect seasonal or future changes that might not match the current labour supply (the possibility of skill mismatches). For instance, as mentioned in Chapter 2, the rapid emergence of modern devices (e.g. computers, smartphones, etc.) have introduced new technologies in the labour market to perform different jobs, such as programmers, data analysts, among others. These accelerated changes have been reflected in the labour demand for skills and have been documented by different authors, such as Acemoglu and Autor (2011). However, the current employment structure might require more time to reflect those changes due to (for instance) the time people need to be trained and offer specific skills.

Considering the advantages and limitations of comparing the vacancy database with household surveys, the following subsection evaluates vacancy data representativeness in Colombia by comparing the vacancy database with Colombian household surveys.

8.3.1. Data representativeness: Vacancy versus household survey information

As mentioned above, the most straightforward way to evaluate vacancy data representativeness is by comparing the results of the occupational structure or employment trends of this source of information with the results from household surveys. The Statistics Office of Colombia (DANE) has carried out a monthly cross-sectional household survey named "*Gran Encuesta Integrada*

de Hogares" (GEIH) since 2006 (see Chapter 3).[122] The GEIH is the main source of official labour market information in Colombia.

8.3.1.1. Occupational structure

At the time this document was written, the DANE was still using the 1970 SOC to classify people's occupations.[123] Perhaps one of the reasons the DANE has not updated their labour supply statistics with ISCO-08 is because the Colombian statistics department still uses manual codifiers (a group of people) to code job titles one by one in its household surveys. As explained in Chapter 6, the manual classification of job titles is a time-consuming task; consequently, updating all household historical records according to ISCO-08 via manual codifiers would require a considerable amount of time and money.

Both manual classification and the use of outdated (and sometimes not well-defined) classifications might be a source of measurement errors. Manual coders might differ from official criteria to classify a job title. Moreover, an outdated classification might not distinguish well some occupational groups. For instance, the 1970 SOC has the following two categories at a two-digit level: code 53 (cooks, waiters, bartenders, and waiters) and code 77 (food preparation workers: bakers, slaughterers, butchers, etc). Consequently, manual coders might not know how to classify a job title such as "Chef" or "Kitchen assistants." The codification might depend on the criteria of each manual coder. In fact, there are codification problems in the GEIH: workers with the same job title (such as "Fried food cook") have different occupational codes (either 53 or 77).

Chapter 6 shows that, despite the relatively large amount of job titles, the Colombian vacancy database is classified automatically using ISCO-08, which is (at this moment) the most up-to-date occupational classification provided by the ILO. Given the advantages of upgrading current labour supply classifications,

---

[122] With a total sample size of approximately 23,000 households monthly, this source of information measures the characteristics of the Colombian workforce. The GEIH collects monthly data representative at national, rural, and urban levels, as well as quarterly data representative at a cities level: Bogotá, Medellín AM, Cali AM, Barranquilla AM, Bucaramanga AM, Manizales AM, Pasto, Pereira, Cúcuta, Ibagué, Montería, Cartagena, Villavicencio, Tunja, Florencia, Popayán, Valledupar, Quibdó, Neiva, Riohacha, Santa Marta, Armenia, and Sincelejo.

[123] This classification was created in 1970 by the Ministry of Labour and Social Protection and the SENA (Cabrera et al. 1997).

the following subsections outline how job titles in the GEIH can be automatically classified according to ISCO-08 to compare the occupational structures of labour supply and demand.

8.3.1.1.1. Categorising the GEIH according to ISCO-08 categories

The GEIH requests the job title for each formal or informal worker. Moreover, all unemployed people are asked about the job position that they are looking for, and unemployed people, who have worked in the past, are asked about their last job position. Consequently, with questions about job titles and the codification of those job titles, it is possible to gather information about occupations for three different groups:

1) Individuals working in formal employment;
2) Unemployed individuals, where occupation refers to the occupation they seek to work in;
3) Individuals working in informal employment.

The procedures described in Section 6.4 were carried out to classify the job titles of the GEIH. Briefly, around 320,000 unique job titles received an occupational code (ISCO-08) by implementing a manual codification, Cascot, and a machine learning algorithm (as described earlier). Once the labour supply information was coded according to ISCO-08, it was possible to carry out the comparison between labour demand and supply information. In total, 419 occupational groups (at a four-digit level) were found in the GEIH.

8.3.1.1.2. Comparing the occupational structures of labour supply and demand

Figure 8.4 shows the percentages of potential job placements (hereafter "job placements" or "job vacancies") from the vacancy database, and the employment level in Colombia from the GEIH (all figures are arranged according to occupational groups at a four-digit ISCO level). Superficially, the chart suggests that a certain level of correlation exists between labour demand and labour supply information. Indeed, the Pearson correlation coefficient is 0.34. Yet, a more

detailed comparison reveals three facts: 1) some occupations do not appear in the vacancy data, but are found in the GEIH data; 2) conversely, occupations that are not listed in the vacancy data do not appear in the labour supply database either; and 3) despite the positive correlation between the occupational structures of labour supply and demand, the vacancy database tends to possess a relatively higher share of technicians and associate professionals and clerical support workers (ISCO major groups 4 and 5), while the GEIH tends to possess a relatively higher share of “Skilled agricultural, forestry, and fishery workers,” “Craft and related trades workers,” “Plant and machine operators and assemblers,” and “Elementary occupations” (ISCO major groups 6, 7, 8 and 9).

Figure 8.4. **Job placements and employment distribution by occupational groups (ISCO-08)**

18.0 %
16.0 %
14.0 %
12.0 %
10.0 %
8.0 %
6.0 %
4.0 %
2.0 %
0.0 %

1114 1324 2112 2149 2261 2353 2433 2619 2653 3122 3154 3312 3342 3432 4212 1313 5120 2512 5321 6221 7131 7233 7411 7532 8131 8160 8331 9214 9612 3433

Job placements
Employed people

Source: Author’s calculations based on GEIH and vacancy information, 2016-2018.

Figure 8.4 evidences, first, that the vacancy database does not contain information about every occupational group in the Colombian economy. Most of the occupations that are not listed in the vacancy database correspond to the military (such as commissioned and non-commissioned armed forces officers, other ranks), agriculture (animal producers, mixed crop and animal producers,

inland and coastal waters fishery workers), or political and social leaders (social welfare managers, senior government officials, etc.). This is understandable given the use of online sources of vacancy information and internet penetration rates in certain zones or sectors of the country (e.g. rural zones). Thus, the vacancy database is not representative for—at least—a significant part of agricultural, government, and armed forces occupations.

Second, the figure shows that occupations that are not listed in the vacancy data do not appear in the labour supply database either, which demonstrates that information from the internet corresponds with—or does not differ from—official national labour market information. For instance, vacancies are not found for nuclear engineers and astronauts, among other occupations, because in Colombia these occupations do not have a market, so there should not be vacancies for these kinds of jobs.[124] This result suggests that online sources of information do not have a surplus of “unreal” or “inappropriate” labour demand in the Colombian context.

Third, as indicated by the figure, the vacancy database has a higher share of “Commercial sales representatives” (ISCO code 3322), “Telephone switchboard operators” (4223), “Stock clerks” (4321), and “Sales and marketing managers” (1221), compared to the GEIH household survey. The high turnover rate of these occupations might explain this issue. Indeed, well-known business platforms such as LinkedIn detail that occupations related to marketing, research, media and communications, as well as support and human resources are amongst those with the highest turnover rates (Booz 2018).

Despite the possibility of higher turnover rates, labour demand (vacancy) and labour supply (household information) display similar patterns. For instance, “Commercial sales representatives” (3322) account for around 15% of the job placements. A similar peak (but of lesser magnitude) is observable in the labour supply information. The same pattern applies for “Accountants” (2411), “Shop sales assistants” (5223), “Sales and marketing managers” (1221), “Mail carriers and sorting clerks” (4412), among others. Consequently, the high job placement share of these occupations is not only due to high turnover rates; these roles

[124] Unless there arises an industry that starts to demand such occupations, in which case there would be no individuals capable of carrying out the tasks required for these new occupations. However, it is not common to observe this phenomenon and this last argument is less plausible given the relatively short period of data collected for this book.

also represent a relatively high portion of Colombian workers. Therefore, the peaks in job placement distribution do not provide strong evidence against data representativeness. On the contrary, this evidence suggests that vacancy data are correlated with labour supply information and some occupations might experience overrepresentation due to higher turnover rates.

In contrast, it is not surprising that the GEIH tends to have a relatively higher share of "Skilled agricultural, forestry, and fishery workers" (ISCO code 6 at one-digit level), as well as "Craft and related trades workers" (ISCO code 7). As mentioned above, low internet access in certain zones or sectors of the country might negatively affect the number of jobs advertised on job portals. Moreover, the GEIH shows a relatively and considerably higher concentration of "Retail and wholesale trade managers" (1420) and "Services managers not classified elsewhere" (1439). A closer look at the job titles demonstrates that these occupations correspond to self-employed people who open and manage their own businesses (for instance, a mini market, a cafeteria, etc.). Consequently, self-employed and "business owner" occupations do not tend to be frequently announced through job portals. In fact, in Colombia, these occupations tend to be found in the informal economy (see Chapter 3). By only considering formal workers in the GEIH, the share of "Retail and wholesale trade managers" (1420) falls to 0.4% and the Pearson correlation coefficient between labour demand and labour supply information increases to 0.39.

The above comparison between labour demand and labour supply information demonstrates at least three facts: 1) The vacancy database is unrepresentative for a considerable proportion of agricultural, government, and armed forces occupations. 2) Despite the high turnover rates of some occupations, labour demand and labour supply demonstrate similar patterns. However, special caution should be taken when analysing occupations with high turnover rates given that this issue might cause an overrepresentation of certain occupational groups. 3) Self-employed and "business owner," as they are informal occupations, are not represented in the vacancy database.

8.3.1.2. Wage distribution of the labour demand and supply

The distribution of wages can be used as an indicator to test the representativeness of the vacancy database. It can be expected that the shapes of the wage

distribution in the vacancy database are similar to the distribution of wages in labour supply data. It is not expected, though, that both the vacancy and the GEIH wages display the same distribution because the vacancy database contains information regarding labour demand and the GEIH survey collects information regarding the supply. Consequently, there are several reasons that might explain the differences between wage distributions in the vacancy database and the GEIH.

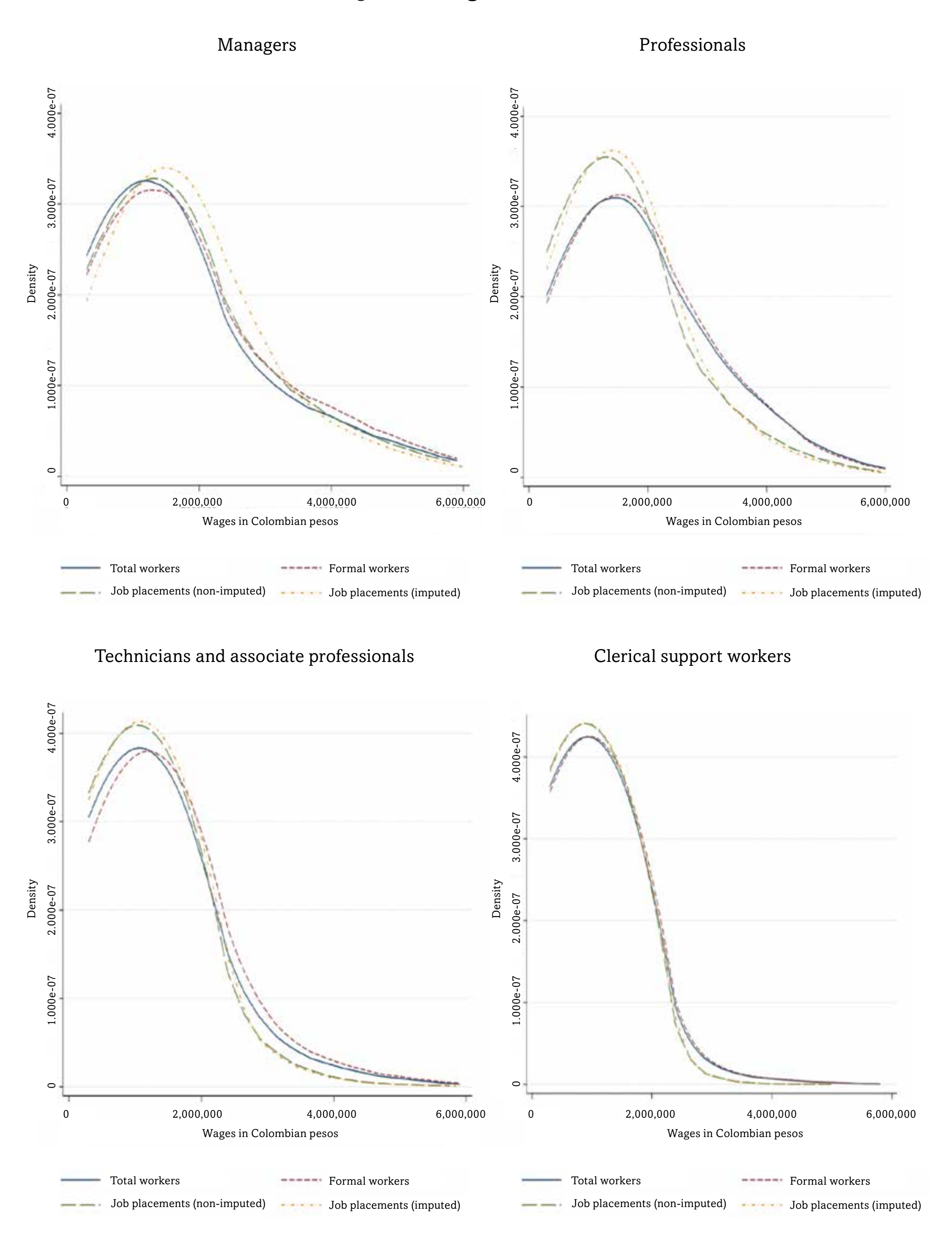
For instance, the vacancy database contains the initial wages that employers are willing to pay for a particular occupation, while the household survey contains a final salary figure, which is agreed upon after a negotiation process between workers and employers. Given this bargaining process, the distribution of wages in the vacancy database for an occupation might be lower than salaries contained in the GEIH. In contrast, skill shortages might explain why the distribution of wages in the vacancy database for an occupation might be higher than wages in the GEIH. However, it is not expected that the bargaining process, skill mismatches, etc., create considerable differences between the shape of distribution in the vacancy and the GEIH datasets.

Similarly, it would be difficult to explain, for example, that for a given occupation wages in the vacancy database are negatively skewed, while the corresponding wages in the GEIH are positively skewed. One possible answer, in this case, is that the labour market is affected by relatively high skill shortage problems, and that, given these mismatches, wage distributions might not display a similar shape. However, and as mentioned above, this argument is not enough to explain the observed differences because if there are such evident and notorious skill shortages, then education and training providers, the government, and, in general, people would have reacted to correct the issue.

Thus, as explained above, Figure 8.5 compares the imputed and non-imputed wage distribution of vacancies (long-dashed and dash-dotted lines, respectively), as well as the wage distribution of total and formal workers in the GEIH (solid and dashed lines, respectively).[125]

[125] Given the large number of occupational groups and the representativeness issues of the GEIH at four digit-level, the graphs are presented at one-digit ISCO.

Figure 8.5. **Wage distributions**



Service and sales workers

Density

0 1.000e-07 2.000e-07 3.000e-07 4.000e-07 5.000e-07

0 2,000,000 4,000,000 6,000,000

Wages in Colombian pesos

Total workers
Formal workers
Job placements (non-imputed)
Job placements (imputed)

Skilled agricultural, forestry and fishery workers

Density

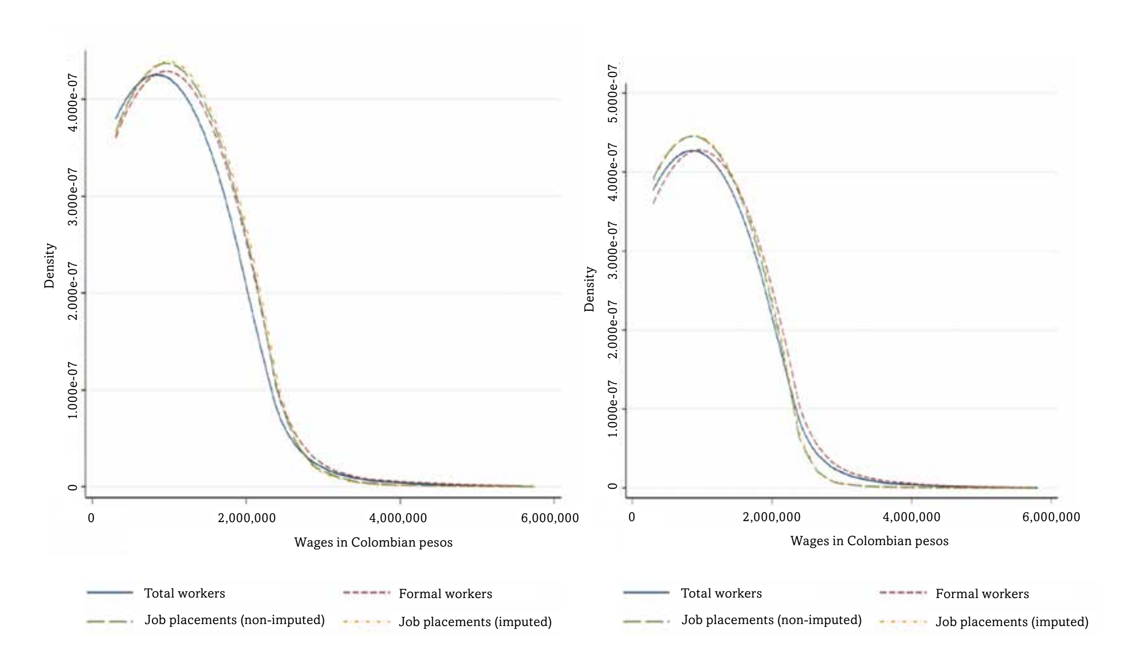
0 1.000e-07 2.000e-07 3.000e-07 4.000e-07

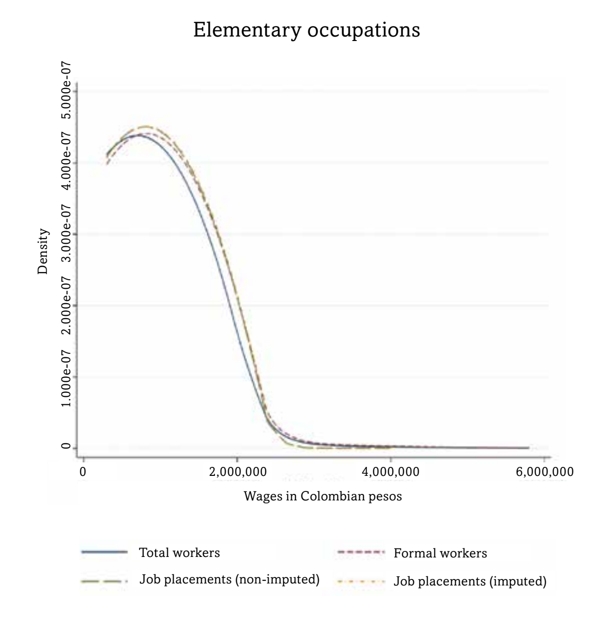
0 2,000,000 4,000,000 6,000,000

Wages in Colombian pesos

Total workers
Formal workers
Job placements (non-imputed)
Job placements (imputed)







Source: Author's calculations based on GEIH and vacancy information, 2016-2018.

The comparison of the distribution of wages, as presented in this figure, reveals four facts. First, regardless of the source of information, high-skilled occupations tend to pay higher salaries than low-skilled occupations. For instance, the median of the wages in the vacancy and the GEIH database for "Managers" are 1,250,000 (non-imputed) Colombian pesos (around £312) per month, 1,614,371 (imputed) pesos (£403), 1,326,000 (total workers) pesos (£331), and 1,500,000 (formal workers) pesos (£375). In contrast, the median of the imputed and non-imputed wages in the vacancy and the GEIH database (total and formal workers) for "Elementary occupations" is 737,700 Colombian pesos (£184) per month. This evidence confirms what is mentioned in the previous section: information regarding human capital demand in Colombia is consistent with information on wages.

Second, worker salaries (GEIH) and job placement wages display a similar shape. Indeed, in most cases, wage distributions almost overlap. This comparison between wage distributions demonstrates that salaries posted on job portals share a similar distribution with wages reported by Colombian workers in the official labour supply survey (GEIH). Moreover, the wage distributions of formal workers are more akin to the distribution of vacancy

wages; for instance, the salary distribution for "Craft and related trades" for the total number of workers is further to the left than for formal workers and for vacancy (job placement) wage distributions. Consequently, the wages of informal workers tend to be lower than formal worker wages and what is offered in vacancies (see Chapter 2).

On the one hand, this evidence suggests that the vacancy database does not contain a considerable number of informal jobs; thus, these data might not be representative for the informal sector. On the other, formal workers and information from job portals (in most cases) have very similar wage distributions. Consequently, the wages in the vacancy database might represent well the "real" salaries that employers are willing to pay for a certain occupation in the formal market, hence information from job portals might consistently represent the "real" distribution of vacancies in Colombia.

Third, despite similarities between vacancies and worker wage distributions, there are some differences too. It is important to note that considerable differences are found in high-skilled occupations: "Managers," "Professionals" and "Technicians and associate professionals." Banfi and Villena-Roldán (2019) found for the Chilean case that companies tend to post explicit wages when experience or educational requirements are relatively low. Consequently, a company's behaviour might affect the vacancy wage distribution of high-skilled occupations. Indeed, imputed vacancy wages tend to be more on the right tail of the distribution than non-imputed vacancy wages. This result suggests that vacancies with inexplicit wages tend to remunerate their workers more than job advertisements with explicit salaries.

Fourth, the fact that job placements and worker wages follow similar distributions suggest that "description bias" might not be a predominant issue. These similarities indicate that worker and vacancy salaries are almost the same, hence there are no particular requirements in the job advertisements (e.g. certifications, use of special technologies, etc.) that might increase or decrease wages in the vacancy data and affect their comparison with wages in the labour supply information.

Alternatively, "description bias" might affect the comparison of the vacancy database with informal jobs. As mentioned above, the wage distribution that considers the total number of workers is more to the left than the one that only considers formal workers. These persistent differences might be explained

by several reasons. One of them is "description bias;" however, even in this scenario, differences between informal wage distributions and the vacancy database (for most occupations) are unremarkable, and the shape of the curves are still similar. Thus, at most, the "description bias" affects vacancy data representativeness for the informal sector.

As mentioned above, the static comparative analysis between job placements and workers is limited, although this analysis allows some occupations to be discarded from the vacancy data, while providing suggestive evidence regarding the representativeness of the other occupational groups. Nevertheless, given the limitations of this analysis more evidence is needed to validate the data representativeness of the vacancy database.

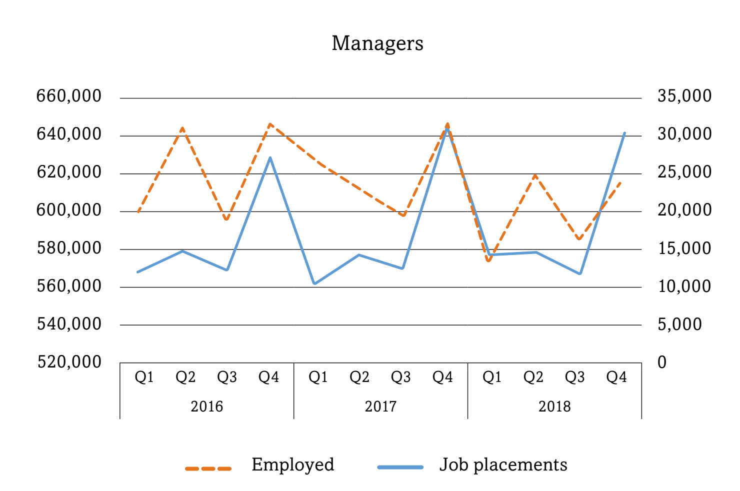
8.3.2. Time series comparison

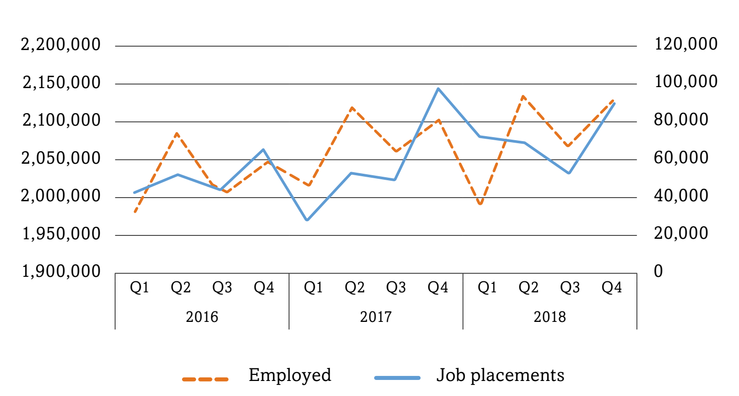
One way to provide further evidence of data representativeness is by comparing labour supply and demand over time ("labour demand and supply series"). It is not expected that this time series follows exactly the same behaviour because some factors (e.g. skill shortage) might affect the correlation between labour demand and labour supply. However, this time series comparison indicates whether economic seasonal and trend effects can be observed in the vacancy database or not. The vacancy database should capture the economic cycles, seasons, and trends to serve as an instrument that informs public policymakers when it is necessary to increase (or decrease) labour supply for specific skills. However, the period covered by the present study is too short to be certain of anything other than seasonal and (short-term) trend effects.

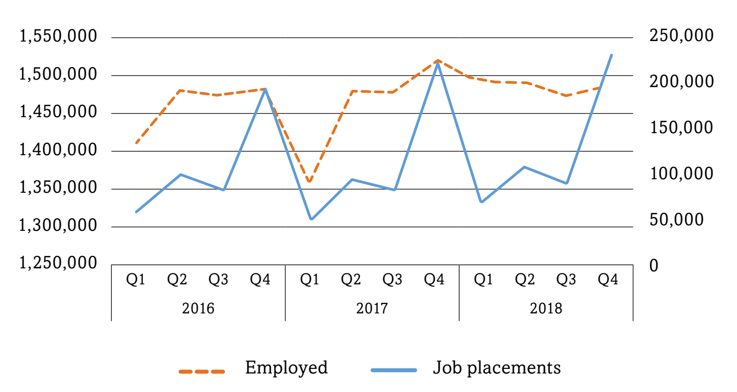
8.3.2.1. Stock of employed people

Figure 8.6 shows the number of vacancies and the number of employed people over time (quarterly from 2016 to 2018) at a one-digit ISCO level (given the large number of occupational groups and the representativeness issues of the GEIH at a four digit-level). The primary axis represents the number of employed people, while the second axis shows the total number of job placements available in a certain quarter from 2016 to 2018.

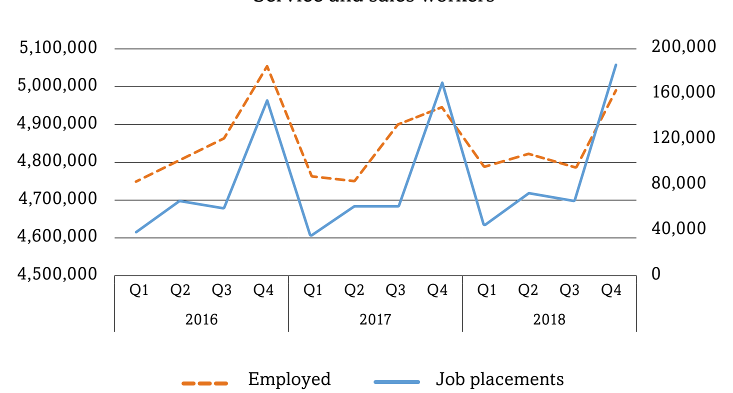
Figure 8.6. **Time series: Total employment and job placements, 2016-2018**

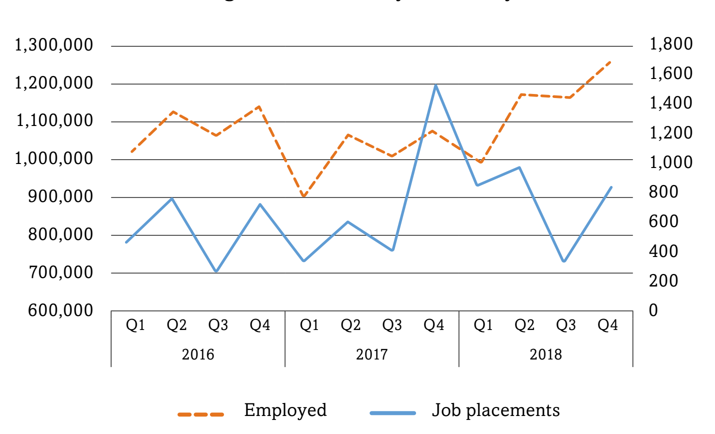


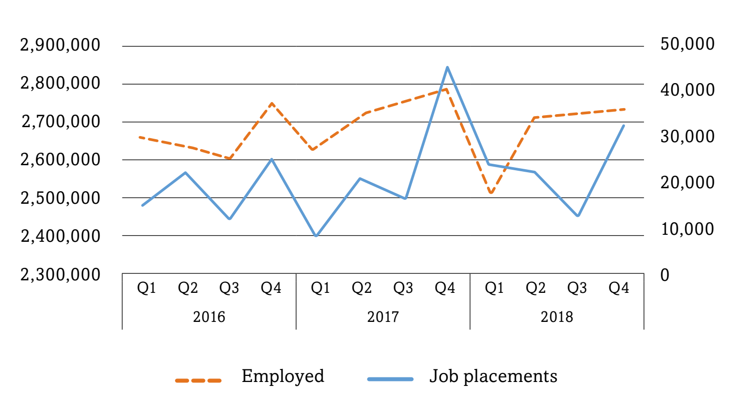


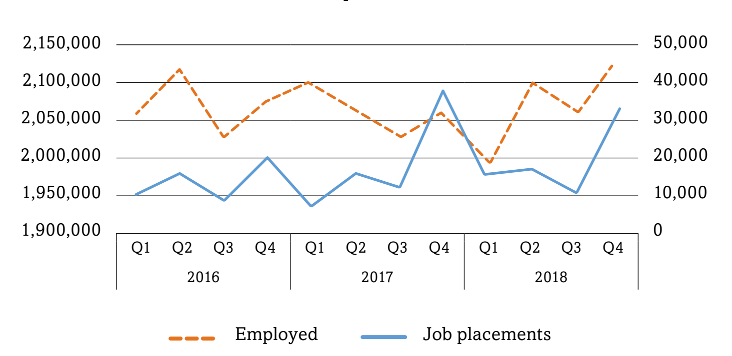


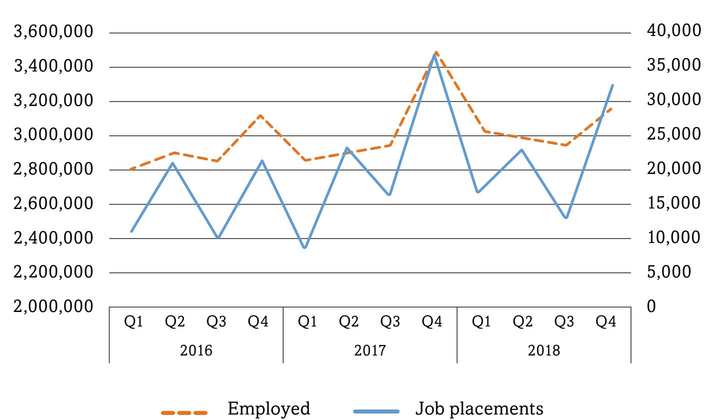

Clerical support workers
860,000
840,000
820,000
800,000
780,000
760,000
740,000
720,000
700,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016 2017 2018
Employed
Job placements


Service and sales workers
5,100,000
5,000,000
4,900,000
4,800,000
4,700,000
4,600,000
4,500,000
200,000
160,000
120,000
80,000
40,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016 2017 2018
Employed
Job placements


Skilled agricultural, forestry and fishery workers
1,300,000
1,200,000
1,100,000
1,000,000
900,000
800,000
700,000
600,000
1,800
1,600
1,400
1,200
1,000
800
600
400
200
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016 2017 2018
Employed
Job placements







Source: Author's calculations based on GEIH and vacancy information, 2016-2018.

As can be observed in the figure, the series of job placements and employed people follows similar economic seasons for all major occupational groups; indeed, even the vacancy database follows similar patterns for "Skilled agricultural, forestry, and fishery workers." Additionally, correlation coefficients range from 0.28 for "Skilled agricultural, forestry, and fishery workers" to 0.87 for "Service and sales workers." This evidence strongly suggests that the vacancy database is a useful instrument to monitor when an occupation is more or less in demand, or when its demand remains unchanged.

Despite the high correlation between the labour demand and supply series, it is still not possible to determine the exact number of vacancies in the Colombian economy; especially, due to the absence of a vacancy census and the issues mentioned in Chapter 4. This limitation might affect the labour market and, specifically, the skill mismatch analysis, because the employment and job placement series might increase at the same time. As the exact number of job placements in the market is unknown, it is not possible to know a priori whether an increase in job placements is going to be compensated for by a rise in the number of workers, or not. In this scenario, it would be difficult to determine skill shortages in the labour market.

However, other information available in the vacancy database or the household survey can dispel any doubts regarding whether there are possible skill mismatches. Perhaps, the most useful variable that can confirm the existence of a skill shortage is the wage variable. As noted in Chapter 2, when a skill mismatch occurs in an occupation or skill, salaries for that segment of the market start increasing. This and the previous subsections prove the consistency of the wage variable and that economic seasons are reflected in the vacancy database. Consequently, when there is an increase in job placements for specific occupations or skills and, in turn, there is an increase in wages, these circumstances strongly suggest the existence of a skill mismatch. Thus, the vacancy database (at this moment) is not able to provide the exact or approximate number of job placements, yet the information can be used to identify possible skill shortages (see Chapter 9).

Moreover, the total number of vacancies in the economy can potentially be estimated. As mentioned in Chapter 2, labour demand is comprised of both the level of employment (satisfied labour demand) and the number of available job vacancies that denote the labour not filled by an employee over a certain

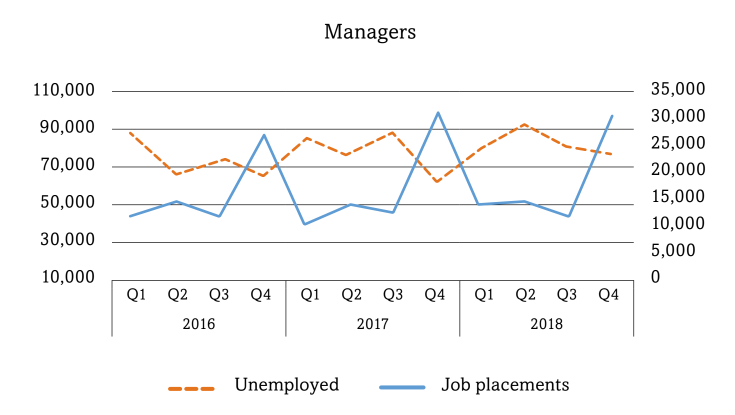
period (unsatisfied labour demand or unmet demand). In turn, the unmet demand is calculated from the separation rate (total number of employees who left their jobs) and the total of new jobs created. By estimating the separation rate, job destruction rate, and sectoral and occupational employment growth rates, similar to Flórez et al. (2017), it might be possible to estimate the level of unmet labour demand and contrast it with the vacancy database. However, the calculation of these parameters will be part of a future work, given the complexity of this task.

8.3.2.2. Stock of unemployed people

The above comparison showed that the vacancy database has a strong correlation with employment rates in Colombia. To provide more evidence regarding the external consistency of the information gathered from job portals, and to demonstrate that vacancy data can be used to build different labour market indicators, this subsection compares the vacancy series with unemployment level. Usually, periods of high unemployment are associated with low levels of vacancies and vice versa (e.g. the Beveridge curve, see Chapter 9).

Figure 8.7 shows a time series to compare unemployment figures against the number of job placements. As expected, in general, these series are negatively correlated for all occupational groups, demonstrating that when there is an increase in the number of job placements, the level of unemployment decreases. Correlation coefficients range from -0.15 for "Service and sales workers" to -0.65 for "Managers." Thus, the results from the vacancy database are consistent with the unemployment series from the official survey. Moreover, these results suggest that it is possible to combine vacancy information with unemployment level to build indicators to monitor the labour market, such as the Beveridge curve, by occupational groups (see Chapter 9).

Figure 8.7. **Time series: Total unemployment and job placements, 2016-2018**



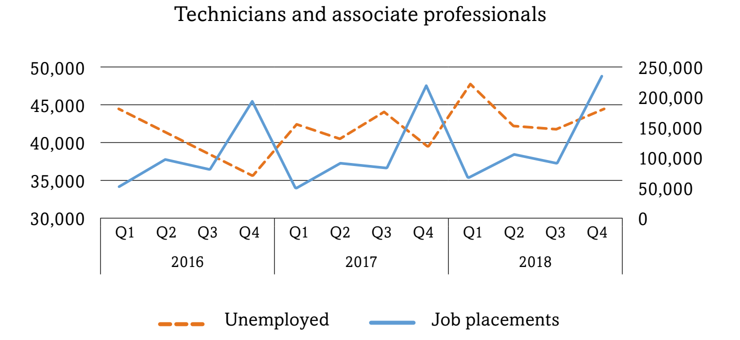
Professionals

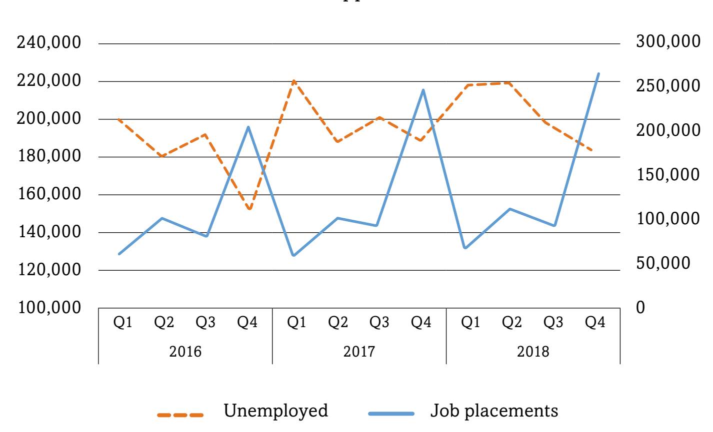
350,000
300,000
250,000
200,000
150,000
100,000

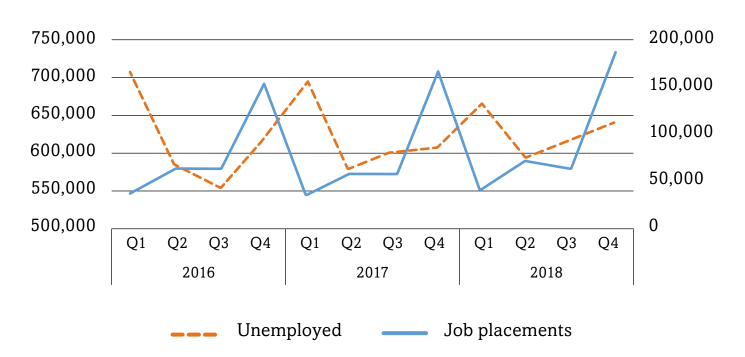
120,000
100,000
80,000
60,000
40,000
20,000
0

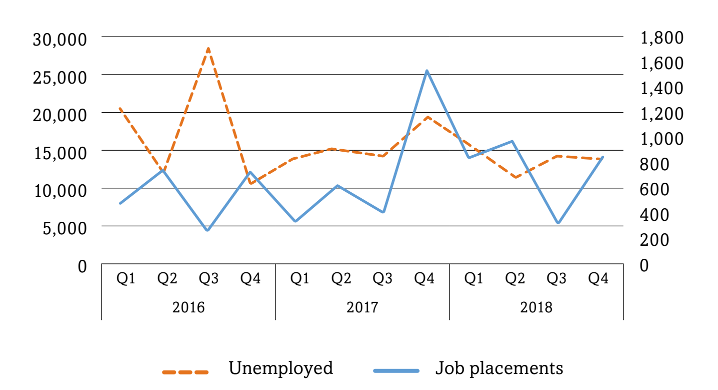
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016 2017 2018

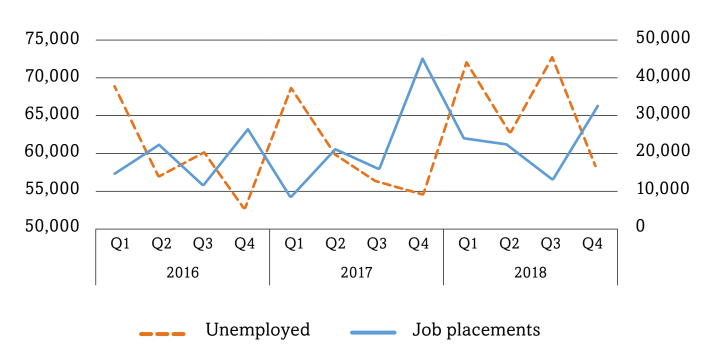
Unemployed
Job placements

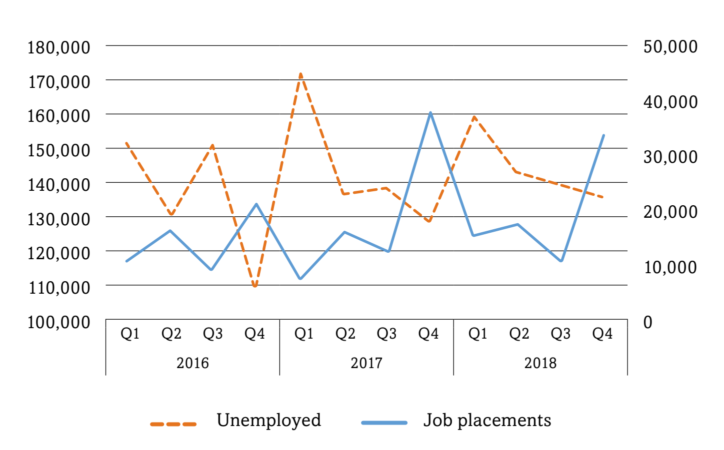


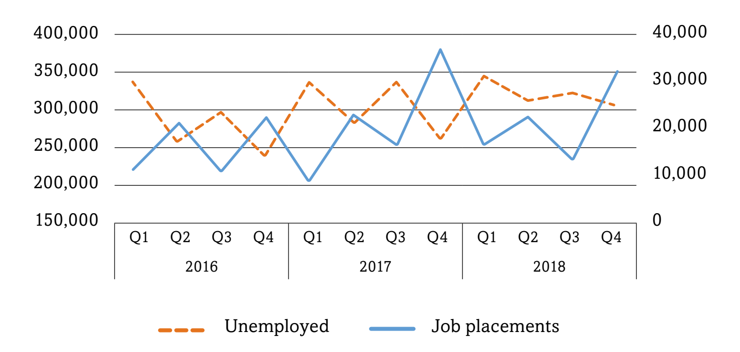

Clerical support workers
240,000
220,000
200,000
180,000
160,000
140,000
120,000
100,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016 2017 2018
Unemployed
Job placements


Service and sales workers
750,000
700,000
650,000
600,000
550,000
500,000
200,000
150,000
100,000
50,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016 2017 2018
Unemployed
Job placements


Skilled agricultural, forestry and fishery workers
30,000
25,000
20,000
15,000
10,000
5,000
0
1,800
1,600
1,400
1,200
1,000
800
600
400
200
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
2016 2017 2018
Unemployed
Job placements







Source: Author's calculations based on GEIH and vacancy information, 2016-2018.

8.3.2.3. New hires (replacement demand and employment growth)

As mentioned above, the comparison between the total workforce and job placements is the most common way to test data representativeness in a vacancy database. However, this exercise might be limited. The total workforce is composed of the total number of employed and unemployed people, while job portals contain information regarding the net and replacement labour demand (see Chapter 2) (LMI for All, n.d.). The total workforce is a measure of the labour market "stock," while the number of job vacancies is a measure of the labour market "flow." Consequently, the similarities (or dissimilarities) between the workforce and job placements time series might be due to other labour dynamics such as participation or dismissal rates, rather than to a causal effect between the number of vacancies and the number of employed or unemployed people.

For instance, the previous subsections showed that positive correlation occurs between the number of job placements and the number of employed people; especially, in the last quarter of each year, the number of employed people and the number of job placements are relatively higher. However, this correlation might be due to a lower dismissal rate. Assuming, at the very least, that the rates of real job openings are consistent in each quarter of the year, it might happen that in the last quarter of the year dismissal rates are relatively lower than in other quarters because employers need to keep more workers for the Christmas season; thus, the number of employed people is higher. Consequently, the vacancy data collected from job portals might not correctly represent the dynamics of real job openings, even when there seems to be a high correlation with the employment and unemployment series.

To test this argument, it is necessary to compare the vacancy series with the net growth,[126] plus replacement demand.[127] It is not possible (so far) in Colombia to identify the total number of vacancies, and much less to distinguish the net growth and replacement demand separately. However, with the Colombian household survey information, it is possible to know when people

[126] Net growth refers to the number of job openings as a consequence of economic growth or decline.

[127] Replacement demand refers to the number of job openings created because of people changing employers, occupations, sector, etc., as well as people temporally leaving their jobs (e.g. sickness), retirement or death.

started to work. Specifically, the GEIH asks the following question: "How long has [interviewed name] been working in this company, business, industry, office, firm or farm continuously?" With this question, it is possible to estimate the number of people who started to work in the previous months (new hires); in other words, the number of new hires (that fill vacancies) created by economic growth (net growth), and the number of vacancies created because people left their jobs (replacement demand). Consequently, new hires have a strong correlation with the number of job openings; thus, if the vacancy database properly represents the dynamics of job openings, the vacancy data should be correlated with the new hires time series.

It is important to note that new hires do not entirely represent labour demand. As mentioned above, the household survey provides information regarding the number of job matches. Consequently, new hires are signified by net growth plus replacement demand matched in the previous months. Nevertheless, there is no strong reason to think that the new hires (matched) time series are not correlated with the number of vacancies available. One argument might be that vacancies occur for certain occupations, but there are no people with the skills and (other) characteristics required. Therefore, vacancies can be created, but not (necessarily) new hires. This argument might be valid for a detailed labour market analysis (e.g. at a four-digit ISCO level). However, general trends and seasonal information for new hires at an aggregated level (e.g. at a one- or two-digit ISCO level) should be reflected in the household survey.

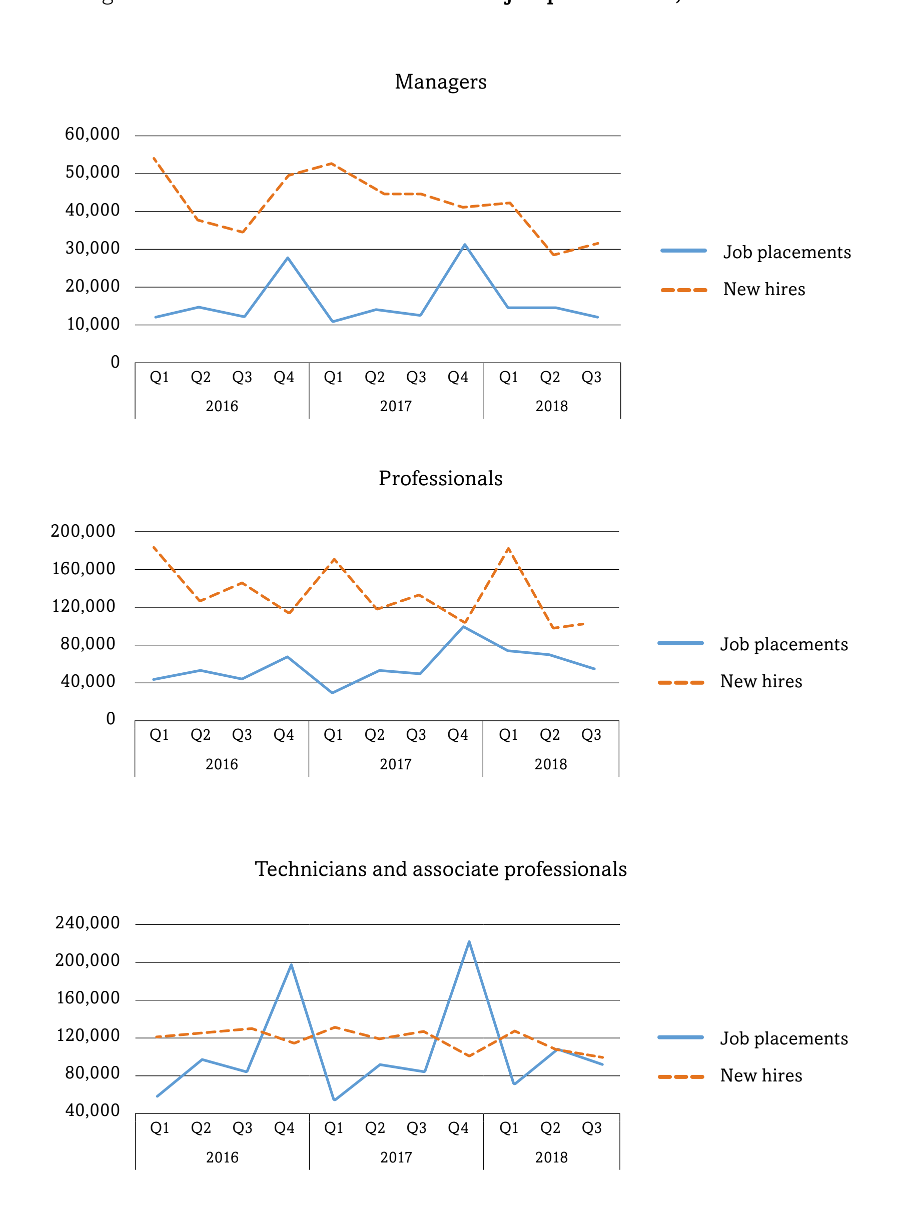
Otherwise, in the Colombian labour market, there are huge barriers such as skill mismatches that prevent people from being hired even when there is an increase in vacancies at the occupational group level (at a one- or two-digit level). Nevertheless, and as mentioned above, this argument does not seem plausible because if there is such an evident barrier to match jobs, the economy and the government would react to correct the issue without the need of a detailed labour market analysis.

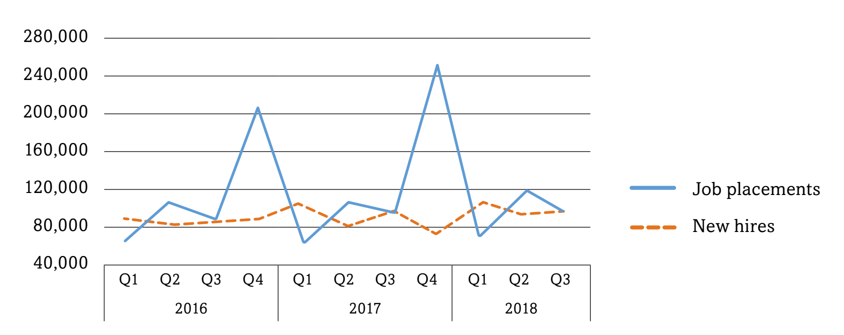
Figure 8.8 depicts the number of new hires and job placements in a quarterly time series.[128] These time series comparisons show an important fact: new hires and job placements have a strong lagged correlation. Indeed, when

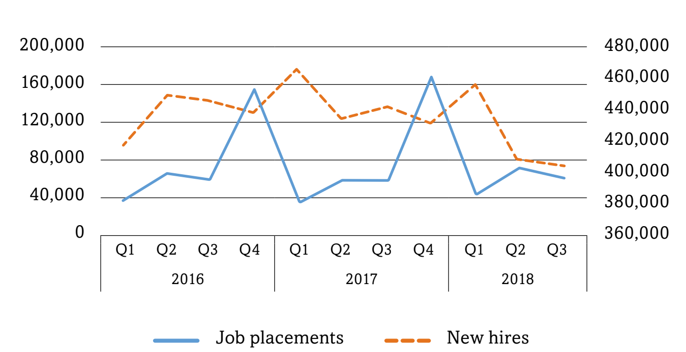
[128] Given the GEIH representativeness issues, the data are quarterly aggregated.

time series are compared within the same period, the Pearson correlation coefficient is between -0.68 and 0.04, and when new hires are lagged by one period (one quarter), the Pearson correlation coefficients sit between 0.17 and 0.70 (except for "Skilled agricultural, forestry, and fishery workers" for which correlation coefficient is -0.01).

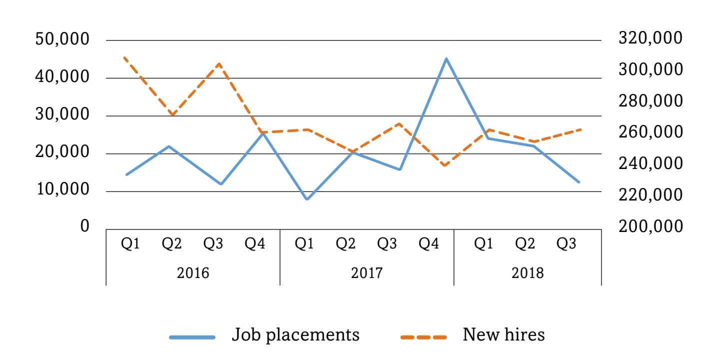
Figure 8.8. **Time series: New hires and job placements, 2016-2018**

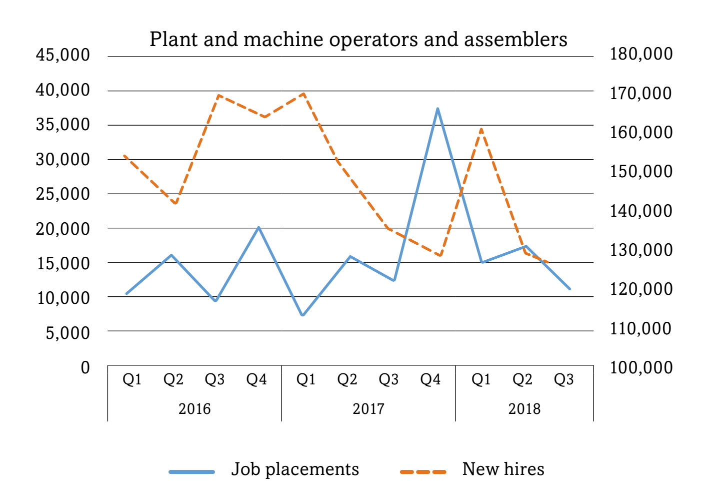


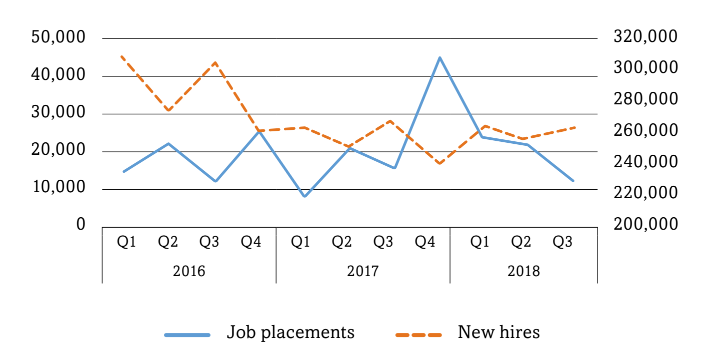

Clerical support workers
280,000
240,000
200,000
160,000
120,000
80,000
40,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Service and sales workers
200,000
160,000
120,000
80,000
40,000
0
480,000
460,000
440,000
420,000
400,000
380,000
360,000
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires


Skilled agricultural, forestry and fishery workers
1,800
1,600
1,400
1,200
1,000
800
600
400
200
0
25,000
20,000
15,000
10,000
5,000
0
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3
2016
2017
2018
Job placements
New hires







Source: Author's calculations based on GEIH and vacancy information, 2016-2018.

The results presented in this figure suggest that there is a lagged effect between the increases and decreases of job placement advertisements and the number of people who occupy these job positions. As mentioned in Chapter 2, posting vacancies is part of the search process, and one of the first steps taken in order to hire workers. Posting the vacancy and hiring the most appropriate worker require time and effort for both the employee and the employers (indeed the median duration of advertising is 1.2 months; see Chapter 7). First, companies need to attract a certain number of workers; after that, companies carry out screening, selecting, and training, among other processes, while workers need to surmount all those processes and, in some cases, work a period of notice with their current employer.

Thus, a lagged correlation is expected between increases/decreases of job advertisements and the moment when people occupy these jobs. Moreover, this lagged correlation shows the dynamics and timing of the hiring process in Colombia. For instance, Chapter 7 showed that for all occupational groups the number of job advertisements sharply increases between October and November, which makes sense given that, as Table 8.4 evidences below, November is the third month when there are additional hires (8.6% of new hires occurred in this month during 2016 and 2018).

However, as also shown in Table 8.4, January is the month in which there are relatively more new hires. This behaviour is because in November companies start hiring people for the Christmas season (see Chapter 7). Nevertheless, in December new hires usually decrease because in this month a considerable portion of people are on vacation (in Colombia, December is well-known as a period where students and most workers take relatively long vacations). Consequently, hiring processes are usually slow in December. On the contrary, January is the start of the new fiscal year when companies become more active again and hire a portion of those people who were contacted and selected in the previous months. This evidence suggests that trends and economic seasons for new hires are strongly correlated with the number of job advertisements, hence the vacancy database adequately represents these trends and the economic season of the total number of job placements.

Consequently, the evidence suggests that for the Colombian case, the vacancy database provides (per se) meaningful information about skills and employer requirements. In general, the occupational structure of the vacancy

Table 8.4. **Monthly distribution of new hires, 2016-2018**

Month	Percent
January	10.40%
August	8.70%
November	8.60%
July	8.50%
September	8.50%
February	8.50%
October	8.40%
June	8.40%
March	8.20%
May	8.10%
April	8.10%
December	5.80%

Source: Author's calculations based on GEIH information, 2016-2018.

information at a four-digit ISCO level is coherent with the information from official surveys, especially, for the urban economy, as well as for formal and non-agricultural occupations. The seasonal and economic trends for a considerable share of the labour market are captured at least at a one-digit ISCO level. Moreover, these data combined with wage, employment, and unemployment information can potentially warn policymakers, educators, and workers about potential skill shortages.

8.4. Conclusion

Any database has limitations. Testing the validity of the database's information is a paramount process to avoid misinterpretation and biases in the analysis. In the case of the vacancy database, which is composed of online job advertisements, different concerns arise (see Chapter 4). For instance, information from the internet might not correlate with general characteristics of the labour market, or the algorithms that collect and organise job advertisements might

fail. Consequently, this chapter has provided an evaluation of the internal and external consistency of the vacancy results.

On the one hand, internal validity refers to the consistency of variables within the vacancy database (Henson 2001; Streiner 2003); this means that the results from a variable in the vacancy database should not contradict the findings from other variables in the same data. The findings of this test show that contradictory or inconsistent results occurring in the Colombian vacancy database are minor, and the magnitude of these measurement errors are insufficient to bias the educational, occupational, sectorial, skills, and wage analyses.

On the other hand, external validity refers to the consistency of the results of the vacancy database when compared with information from other sources (in other words, data representativeness) (Rasmussen 2008; Stopher 2012). The vacancy data per se can provide valuable answers about what people should be trained in at a low cost (time and money). Nevertheless, testing the data "selection bias" of the vacancy database is challenging because of the absence of a vacancy census, or any official data that supply the total number of vacancies in Colombia (statistical universe).

Despite different difficulties, this chapter has provided an external evaluation utilising sources of information available in the country. Thus, a static comparison was made between labour supply and vacancy information. First, the occupational structure of the vacancy database (labour demand) and the GEIH (labour supply) was compared. This comparison provided three conclusions: 1) the vacancy database is not representative for a considerable part of agricultural, government, and armed forces occupations; 2) particular caution should be taken when analysing occupations with high turnover rates as this issue might cause an overrepresentation of specific occupational groups; and 3) self-employed individuals ("business owners") and informal occupations are not represented in the vacancy database. This evidence suggests that the vacancy database better represents the formal and urban labour market in Colombia.

Second, a comparison between the distribution of wages in the vacancy database and the GEIH was carried out. This exercise suggests that wages in the vacancy database well represent the "real" salaries that employers are willing to pay for a particular occupation. The comparison also shows that the vacancy database might consistently represent the distribution of vacancies in Colombia.

Moreover, the vacancy database should capture economic seasons, cycles, and trends to serve as an instrument that can inform public policymakers when it is necessary to increase (or decrease) labour supply for specific skills. Consequently, time series comparisons between the number of vacancies and employed and unemployed people, as well as new hires were made to establish whether economic seasons could be observed in the vacancy database or not. This comparison showed that information from job portals captures and represents economic seasons in Colombia. In general, when the level of job placements increases, so does the level of employment; conversely, when there is an increase in the number of job placements, the level of unemployment decreases. Importantly, the comparison between new hires and job placements revealed that economic trends and seasons for new hires are strongly (lagged) correlated with the number of job advertisements, hence the vacancy database adequately represents the "real" trends and economic seasons of the total number of job placements. Thus, training providers could potentially use the vacancy database information to estimate when training provision should be increased, decreased or maintained. However, so far, it was not possible to analyse economic cycles due to the relatively short period of information available from the database (three years).

It is not possible either (at this moment) to determine the exact number of vacancies in the Colombian economy, mainly, because of the absence of a vacancy census. However, it is not necessary to include a precise amount of vacancies in the economy to identify possible skill shortages, among other essential characteristics of the labour market. A rigorous analysis using information from online job portal vacancies and GEIH data (such as wages, trends, occupational structure, etc.) provide sufficient information to design indicators (such as the Beveridge curve or wage and employment trends) and determine possible skill shortages for a segment of the Colombian labour market.

Thus, the vacancy database, in general, is representative of a considerable set of formal, non-agricultural, non-governmental, non-military, and non-self-employed ("business owners") occupations between 2016 and 2018. Despite the fact that the vacancy information does not capture a considerable share of agricultural jobs, the relatively few observations in the vacancy database for those occupations might provide insights for policymakers, educators, and workers about new skill requirements and general trends for some agricultural occupations.

9. Possible Uses of Labour Demand and Supply Information to Reduce Skill Mismatches

9.1. Introduction

As explained in Chapters 3 and 4, Colombia does not have a proper system to identify possible skill mismatches (skill shortages), hence education and training providers experience difficulties in training people according to current employer requirements. As a potential solution to this issue, Chapters 7 and 8 have demonstrated that job portals are rich sources of representative information for the analysis of a considerable segment of the Colombian labour demand (job openings). The systematic collection and depuration of this information via web scraping and text mining, among other techniques, provide (at a low cost) valuable information about skill requirements that employers demand, and the structure and trends of this labour demand. Consequently, this novel source of vacancy information is useful for reducing imperfect information issues and, more specifically, for tackling two main issues in the Colombian labour market: unemployment and informality. Thus, this chapter shows how the vacancy database, along with household survey information, can be used as a tool to address the labour market issues mentioned above.

Given that the occupational structure of the database, as well as seasonal and other vacancy information trends are broadly consistent with results from official surveys, this indicates three advantages of the vacancy database. First, the vacancy database can be used to describe the main characteristics of unmet labour demand (e.g. occupational structure, wages, educational requirements, etc.), as well as its structure and changes over time. Vacancy information combined with labour supply information generates the possibility of describing and comparing the characteristics of labour demand and supply in Colombia, while a descriptive analysis provides an understanding of the structure of the Colombian labour market and labour market issues; for example, where possible or more remarkable skill shortage problems occur.

Second, and more importantly, with the combined use of household surveys (GEIH) and the vacancy database, a set of macro-indicators are proposed to identify current skill shortages. For instance, the existence of a skill mismatch is suggested when there is an increase in job placements for specific occupations or skills and, in turn, there is an increase in real wages. In addition, when there is an increase in the unemployment rate and a decline of job placements and real wages for a certain occupation, these features also suggest the existence of a skill mismatch.

Third, as shown in Chapters 7 and 8, vacancy information provides detailed and updated information regarding employer requirements at a low cost and in real time. Specifically, the vacancy information provides insight about new job titles and skills demanded in Colombia; consequently, job portals are a valuable source of information to keep occupational classifications updated and monitor composition and skill trends by occupation. With the regular updating of occupational classifications, education and training providers have useful inputs as a basis for their curricula (according to employer requirements), and public policymakers can identify any barriers (or lack of skills) that obstruct the entrance of people into formal economy.

Given these three advantages of the information collected from job portals, this chapter discusses how the vacancy database can be used to build a detection system of skill shortages, and to regularly update occupational classifications according to employer requirements. The second section of this chapter characterises the labour market (formally and informally employed, as well as unemployed) by educational and occupational levels from 2016 to 2018. The third section elaborates on a set of macro-indicators—for the first time in Colombia—within the vacancy database's labour demand and supply information for the identification of possible skill shortages. Finally, the fourth section illustrates how detailed information from vacancies (job descriptions) can be used to update occupational classifications (ISCO) and labour force skills according to employer requirements.

9.2. Labour market description

The theoretical framework of this book (see Chapter 2) has stressed that a considerable proportion of unemployment and informal economy is explained by a misallocation between skills possessed by job seekers and skills demanded by employers. Moreover, it has been argued that wages in the formal economy tend to be higher than in the informal economy; thus, informal workers have incentives to be part of the formal economy. Indeed, Chapter 3 has shown that the Colombian labour market is characterised by prolonged and relatively high unemployment and informality rates (in 2017, around 47% of workers were informal, and the unemployment rate was approximately 10%), and informal workers earn between 40% and 60% less than their formal peers. Additionally, the evidence suggests that one of the leading causes of unemployment in Colombia is due to skill mismatches between labour demand and supply.

This section describes, by occupation, the characteristics of formally and informally employed workers, and those who were unemployed, from 2016 to 2018.[129] This characterisation of the labour market illustrates the structure of the Colombian labour market and provides an idea of labour market issues, for example, where possible or more remarkable skill mismatch problems occur. One of the most distinctive elements of this characterisation is that it shows—for the first time—a disaggregated occupational analysis based on the Colombian household survey using a relatively updated classification such as ISCO-08. As shown in the previous chapter, one of the most important advantages of reclassifying the household survey according to ISCO-08 is that this classification allows comparisons with labour demand information—and, in further research, it will enable international comparisons. Perhaps the reason why researchers had not considered using the occupational variable before for identifying skill mismatches is that this variable was aggregated and outdated given that updating all household historical survey records according to ISCO-08 via manual codifiers would require a considerable amount of time and money (Chapter 8). However, the previous chapters have shown that it is

129 For the employment time series analysis, data were available from 2010 to 2018.

possible to overcome these issues with the help of tools such as Cascot and machine learning techniques.

As mentioned in Chapter 8, official labour market information (GEIH) is representative of urban and rural areas, while the vacancy information might not provide accurate results for the country's rural zones. This chapter examines the results from the GEIH regarding Colombian urban zones to make adequate comparisons between the labour supply and labour demand information.

9.2.1. Colombian labour force distribution by occupational groups

Tables 9.1 and 9.2 describe the occupational composition of formal and informal workers and unemployed people at a four-digit ISCO level[130] from 2016 to 2018 (the full tables can be found in Appendix H, Tables H.1 to H.3). Most of the formal workers are "Sales demonstrators," followed by "(Secondary or university) education teachers",[131] and "Security guards," while most of the informal workers are "Sales demonstrators," "Domestic cleaners and helpers," and "Car, taxi, and van drivers."

Table 9.1. **Occupational distribution of Colombian workers**[132]

#	ISCO title	Formal workers	ISCO title	Informal workers
1	Sales demonstrators	4.8%	Sales demonstrators	16.4%
2	(Secondary or university) education teachers	4.5%	Domestic cleaners and helpers	6.0%
3	Security guards	3.7%	Car, taxi, and van drivers	6.0%
4	Cleaners and helpers in offices, hotels, and other establishments	3.6%	Stall and market salespersons	3.7%

[130] Given that the GEIH might have representativeness issues when the data is disaggregated at a four-digit ISCO level, the results at a four-digit level are indicative of the general structure of the Colombian labour market but they might not exactly represent the distribution of the labour force by occupational groups.

[131] In most cases, information available in the GEIH does not distinguish between primary, secondary and university teachers.

[132] Occupations with the lowest frequency (10% of occupations in the GEIH) were dropped to avoid representativeness issues and outliers.

#	ISCO title	Formal workers	ISCO title	Informal workers
5	Car, taxi, and van drivers	3.0%	Cleaners and helpers in office, hotels, and other establishments	3.3%
6	Stock clerks	2.0%	Cooks	2.9%
7	Health care assistants	1.9%	Commercial sales representatives	2.3%
8	Building and related electricians	1.8%	Bricklayers and related workers	2.1%
9	Accounting and bookkeeping clerks	1.7%	Child care workers	2.1%
10	Waiters	1.5%	Building and related electricians	1.9%
11	Welders and flame cutters	1.5%	Beauticians and related workers	1.9%
12	Primary school teachers	1.5%	Sewing machine operators	1.9%
13	Child care workers	1.5%	Services managers not elsewhere classified	1.8%
14	Sewing machine operators	1.4%	Shop keepers	1.8%
15	Mail carriers and sorting clerks	1.3%	Kitchen helpers	1.7%
16	Cooks	1.3%	Motorcycle drivers	1.6%
17	Cashiers and ticket clerks	1.3%	Motor vehicle mechanics and repairers	1.6%
18	Contact centre information clerks	1.1%	Construction supervisors	1.4%
19	Kitchen helpers	1.0%	Freight handlers	1.2%
20	Senior officials of special-interest organisations	1.0%	Waiters	1.2%

Source: Author's calculations based on GEIH information, 2016-2018.

According to Table 9.2,[133] most unemployed people in Colombia are seeking jobs as "Sales demonstrators," "Cleaners and helpers in offices, hotels, and other establishments," and "Domestic cleaners and helpers."

[133] As mentioned in Chapter 8, the GEIH asks unemployed people: "What kind of job (occupation) are you looking for?" This question identifies what occupations unemployed people are trying to find.

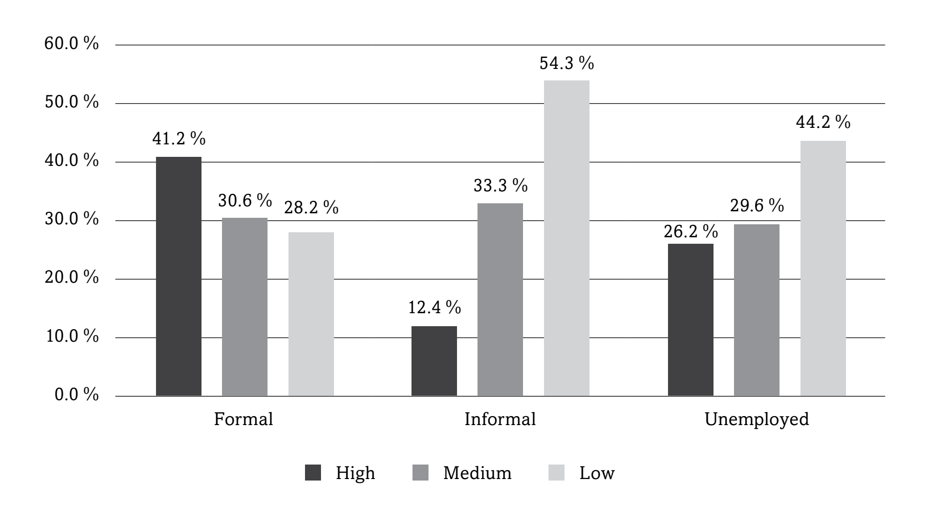
Table 9.2. **Occupational distribution of jobs sought by unemployed people in Colombia**

#	ISCO title	Unemployed
1	Sales demonstrators	13.9%
2	Cleaners and helpers in offices, hotels, and other establishments	4.9%
3	Domestic cleaners and helpers	4.4%
4	Building and related electricians	3.2%
5	Waiters	3.1%
6	Security guards	3.1%
7	Stock clerks	2.7%
8	Car, taxi, and van drivers	2.7%
9	Health care assistants	2.0%
10	Accounting and bookkeeping clerks	2.0%
11	(Secondary or university) education teachers	2.0%
12	Administrative and executive secretaries	1.7%
13	Kitchen helpers	1.6%
14	Contact centre information clerks	1.6%
15	Cooks	1.6%
16	Cashiers and ticket clerks	1.5%
17	Bricklayers and related workers	1.5%
18	Sewing machine operators	1.4%
19	Child care workers	1.2%
20	Construction supervisors	1.1%

Source: Author's calculations based on GEIH information, 2016-2018.

Figure 9.1 summarises the labour market structure of the Colombian workforce by occupational groups: 41.2% of the formal workers are in high-skilled occupations, followed by medium- and low-skilled occupations at 30.6% and 28.2%, respectively. Conversely, low-skilled occupations represent 54.3% of the informal workers and 44.2% of those unemployed. This evidence seems to confirm what was mentioned in Chapter 3, that a lack of skills is a prevalent problem in Colombia and contributes to high rates of unemployment and informality.

Figure 9.1. **Occupational distribution of the Colombian workforce by skill level**



Source: Author's calculations based on GEIH information, 2016-2018.

9.2.2. Unemployment and informality rates

The above results show the composition of the Colombian workforce by occupational groups, and they allow identifying the general structure and patterns in the labour force by occupation, skill level, and formal/informal/unemployed workers. However, the above analysis does not indicate which occupational groups tend to have the highest informality and unemployment rates. For instance, Table 9.1 shows that 16.4% of the informal workers are "Sales demonstrators." The high proportion of this occupation in the informal labour market might be due to the fact that a considerable number of Colombian workers is employed in this occupation. It might well be that they have a low informality rate because the number of formal sales demonstrators far exceeds the number of informal sales demonstrators.

It is essential to observe these rates because they demonstrate which occupational groups tend to be more/less exposed to unemployment or informality. Consequently, Table 9.3 shows that the occupations with the highest informality rates are "Domestic cleaners and helpers," "Motorcycle drivers," and "Shop keepers" (full tables can be found in Appendix H, Tables H.4 and H.5).

Table 9.3. **Occupations with higher informality rates**

#	ISCO title	Informality rate
1	Domestic cleaners and helpers	99.8%
2	Motorcycle drivers	99.0%
3	Shop keepers	97.3%
4	Tailors, dressmakers, furriers, and hatters	96.7%
5	Street food salespersons	96.6%
6	Stall and market salespersons	95.3%
7	Sewing, embroidery, and related workers	94.1%
8	Drivers of animal-drawn vehicles and machinery	93.6%
9	Potters and related workers	92.3%
10	Clearing and forwarding agents	92.2%
11	Sales workers not elsewhere classified	92.0%
12	Beauticians and related workers	90.7%
13	Handicraft workers in textile, leather, and related materials	90.7%
14	Hairdressers	89.2%
15	Bicycle and related repairers	89.0%
16	Fast food preparers	87.6%
17	Laundry machine operators	87.2%
18	Refuse sorters	86.0%
19	Street vendors (excluding food)	84.9%
20	Bricklayers and related workers	83.7%

Source: Author's calculations based on GEIH information, 2016-2018.

By contrast, Table 9.4 presents occupations that tend to have lower informality rates, where "Computer network professionals," "Dieticians and nutritionists," and "Geologists and geophysicists" are among the occupations with the lowest level of informality.

Table 9.4. **Occupations with lower informality rates**

#	ISCO title	Informality rate
1	Computer network professionals	0.0%
2	Dieticians and nutritionists	0.3%
3	Geologists and geophysicists	0.9%

#	ISCO title	Informality rate
4	Computer network and systems technicians	1.2%
5	Mathematicians, actuaries, and statisticians	1.2%
6	Psychologists	1.5%
7	Metal production process controllers	1.6%
8	Mining supervisors	1.8%
9	Travel attendants and travel stewards	1.9%
10	Legislators	1.9%
11	Vocational education teachers	2.0%
12	Software developers	2.1%
13	Sweepers and related labourers	2.4%
14	University and higher education teachers	2.5%
15	Visual artists	2.6%
16	Filing and copying clerks	2.6%
17	Secondary education teachers	2.7%
18	Health services managers	2.7%
19	Statistical, finance, and insurance clerks	2.8%
20	Economists	2.8%

Source: Author's calculations based on GEIH information, 2016-2018.

Additionally, using the vacancy database information, it is possible to identify the skills demanded by occupations with low informality rates. For instance, for "Computer network professionals," the most required skills are APL (A Programming Language), customer service, communication, and knowledge in alarm and control systems. Consequently, these low rates, along with vacancy skills information, might suggest what occupations and specific skills people should possess to improve their probabilities of finding a formal job. However, as will be discussed in the following section, there are other variables to consider before determining a skill shortage in this way.

Based on information about what jobs are being sought by potential workers, Table 9.5 presents occupations with a higher unemployment rate. As evidenced in the table, "Environmental engineers" have the highest unemployment rate (36.7%), followed by "Geologists and geophysicists" (26.1%) and "Sociologists, anthropologists, and related professionals" (25.4%). Additionally, also shown in this table, occupations with higher unemployment rates tend to have a

prolonged (above average) duration of unemployment. These results do not contradict the unemployment rates reported by the DANE: according to the Colombian office of national statistics, the unemployment rate for undergraduates was relatively high (around 10%) in 2016,[134] and the average duration of unemployment for undergraduates is 26 weeks, while it is 18 weeks for people with only a high school certificate.

Table 9.5. **Occupations with higher unemployment rates**

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
1	Environmental engineers	36.7%	29.3
2	Geologists and geophysicists	26.1%	31.7
3	Sociologists, anthropologists, and related professionals	25.4%	24.8
4	Economists	22.7%	46.3
5	Philosophers, historians, and political scientists	22.7%	40.3
6	Survey and market research interviewers	22.5%	21.0
7	Contact centre information clerks	22.1%	18.1
8	Filing and copying clerks	21.8%	25.9
9	Veterinary technicians and assistants	21.6%	10.8
10	Environmental and occupational health inspectors and associates	20.7%	27.9
11	Enquiry clerks	20.0%	27.9
12	Mining engineers, metallurgists, and related professionals	19.9%	33.1
13	Receptionists (general)	19.2%	26.1
14	Stock clerks	18.8%	18.6
15	Mechanical engineers	18.7%	25.9
16	Sports, recreation, and cultural centre managers	18.5%	12.9
17	Business services agents not elsewhere classified	18.4%	20.8
18	Social work and counselling professionals	17.9%	29.3
19	Information and communications technology operations technicians	17.5%	24.9
20	Psychologists	17.1%	29.4

Source: Author's calculations based on GEIH information, 2016-2018.

[134] See https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/fuerza-laboral-y-educacion.

Tables 9.4 and 9.5 show the importance of analysing unemployment and informality rates at the same time. Occupations such as "Geologists and geophysicists," "Economists," or "Filing and copying clerks" tend to have low informality rates, but high unemployment rates and prolonged unemployment periods. Consequently, an increase in labour supply in occupations with relatively low informality rates might increase the unemployment rate. Thus, any public policy that attempts to reorient labour supply according to employer requirements should consider unemployment and informality rates together.

By contrast, Table 9.6 presents occupations with lower unemployment rates. The data presented in the table show that "Religious professionals" have the lowest unemployment rate (0.3%), followed by "Motorcycle drivers" (0.5%) and "Shopkeepers" (0.7%). Moreover, as also evidenced in this table, occupations with lower unemployment rates tend to have a shorter (below average) duration of unemployment.

Table 9.6. **Occupations with lower unemployment rates**

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
1	Religious professionals	0.3%	19.3
2	Motorcycle drivers	0.5%	8.6
3	Shop keepers	0.7%	16.9
4	Bicycle and related repairers	0.9%	13.7
5	Legislators	0.9%	16.3
6	Tailors, dressmakers, furriers, and hatters	1.0%	23.6
7	Potters and related workers	1.0%	8.5
8	Handicraft workers in textile, leather, and related materials	1.1%	24.3
9	Pawnbrokers and money-lenders	1.1%	6.0
10	Dairy-products makers	1.3%	15.2
11	Stall and market salespersons	1.4%	20.7
12	Weaving and knitting machine operators	1.4%	26.0
13	Sewing, embroidery, and related workers	1.4%	24.2
14	Debt-collectors and related workers	1.5%	13.1
15	Education managers	1.6%	36.3

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
16	Refuse sorters	1.6%	3.3
17	Travel consultants and clerks	1.8%	18.0
18	Contact centre salespersons	1.9%	19.0
19	Accounting associate professionals	1.9%	17.8
20	Hairdressers	1.9%	19.4

Source: Author's calculations based on GEIH information, 2016-2018.

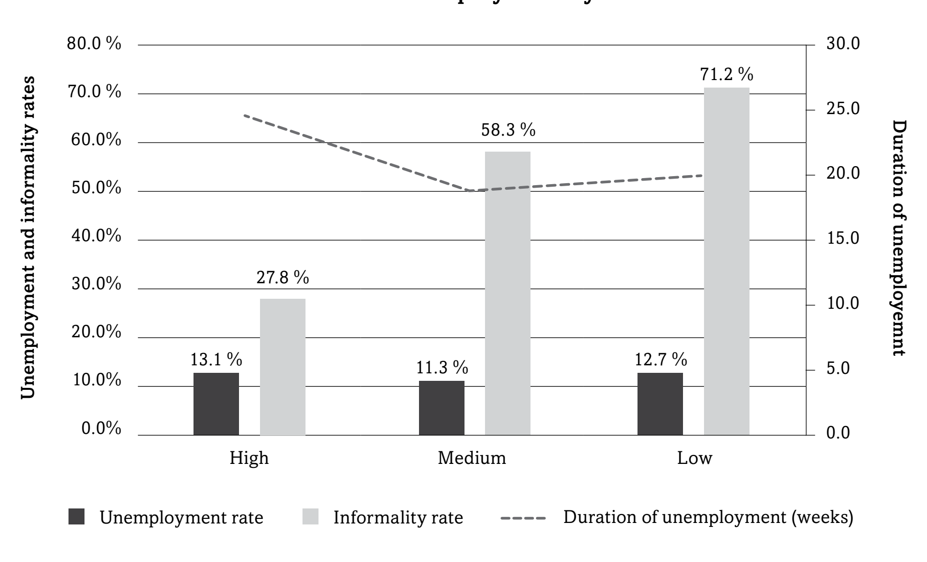
Additionally, the results from Table 9.6 can be complemented with vacancy database information. For instance, for "Motorcycle drivers," the most demanded skills are customer service, sales activities, work in an organised manner, and count money (see Section 9.4).

Importantly, Tables 9.3 and 9.6 also highlight the importance of analysing unemployment and informality rates at the same time in order to draw proper public policy advice from data. Occupations such as "Motorcycle drivers," "Shopkeepers," "Refuse sorters," and "Hairdressers," among others, tend to have low unemployment rates and shorter unemployment periods, but high informality rates. Consequently, increased labour supply in occupations with relatively low unemployment rates might increase the informality rate.

Figure 9.2 summarises labour informality and unemployment rates by occupation skill level. Low-skilled occupations have an informality rate of 71.2%, followed by medium- and high-skilled occupations with 58.3% and 27.8%, respectively. In contrast, high- and low-skilled occupations reported the highest unemployment rates, with 13.1% and 12.7%, respectively. Moreover, as also shown in this figure, the duration of unemployment is higher for high-skilled people.

According to the theoretical framework of this book (see Chapter 2) and the evidence presented in this chapter, skill mismatches are widespread in the Colombian economy, the consequences of which are reflected in its relatively high unemployment and informality rates. However, low-skilled occupations tend to present more signs of oversupply (high informality and unemployment rates). Consequently, Colombian public policies should pay special attention to informing, educating, and training people with low skills according to the employers' needs.

Figure 9.2. **Unemployment and informality rates and duration of unemployment by skill level**

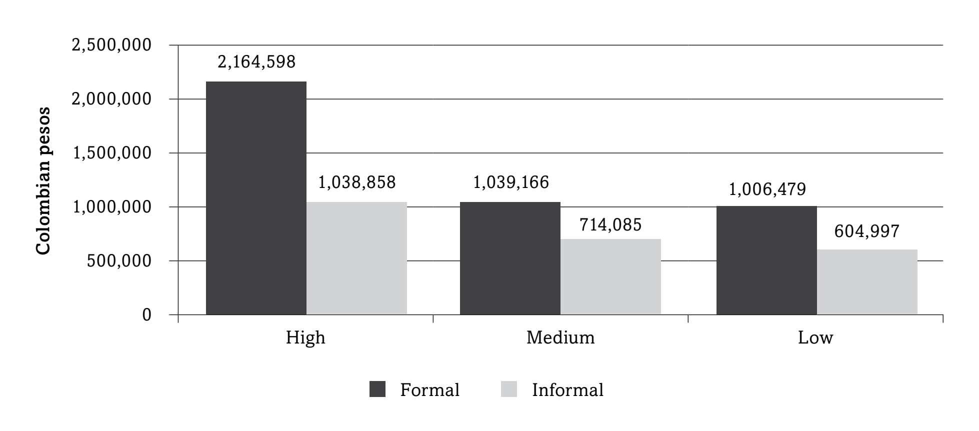


Source: Author's calculations based on GEIH information, 2016-2018.

As mentioned in Chapters 2 and 3, the informal economy, overall, tends to pay lower salaries than the formal economy. Figure 9.3 shows the average wages of formal and informal workers by skill level. As can be observed, there is a considerable wage gap between formal and informal workers across all skill groups. However, the difference between formal and informal high-skilled workers is higher: formal workers in high-skilled occupations earn 52.0% more than their informal peers. Furthermore, formal low- and medium-skilled workers earn 39.9% and 31.3% more than their informal peers, respectively. Thus, this supports the claim indicated in Chapter 2 that informal workers (in terms of income) have an incentive to be part of the formal economy.

In summary, the informality and unemployment rates in Colombia are relatively high. Informal labour (once compared with formal and unemployed population) is mainly composed of adults (more than 29 years old) with a high school educational level or less (see Chapter 3). On the one hand, in concordance with previous results, people in low-skilled occupations have the highest informality rates, while, on the other hand, the unemployed population is mainly composed of young adults (less than 29 years old) (see Chapter 3). Moreover, people in high- and low-skilled occupations have the highest unemployment rates and

Figure 9.3. **Average wages of formal and informal workers by skill level**



Source: Author's calculations based on GEIH information, 2016-2018.

prolonged unemployment periods. Consequently, the evidence suggests that informality issues tend to occur more frequently for adults with (at most) high school education, who work in low-skilled occupations, while unemployment issues occur more frequently in groups of people who are less than 29 years old and work in low- or high-skilled occupations. Thus, regardless of the skill group, the Colombian labour market displays potential signals of skill mismatches.

Nevertheless, low-skilled occupations tend to express more signs of over-supply: 1) a considerably higher informality rate compared to medium- and high-skilled occupations; and 2) a high unemployment rate (slightly lower than the high-skilled unemployment rate). These results suggest that, in Colombia, skill shortages might be more frequent in medium- and high-skilled occupations (see Section 9.3).

Differences in the average wages of formal and informal workers by skill level show that informal and unemployed workers—regardless of their skill level—have a strong incentive to be part of the formal economy. As explained in Chapters 2 and 3, despite this financial incentive to participate in the formal economy, the evidence suggests that misallocation between skills possessed by job seekers and skills demanded by employers makes the formalisation of a considerable part of the Colombian economy a challenge. Thus, policymakers in Colombia need to administer a national and systematic analysis of human resources demand and supply, and act based on reliable data to tackle high unemployment and informality rates, especially, in low-skilled occupations.

Moreover, for a proper human resources analysis, it is necessary to consider and compare occupational unemployment and informality rates. Some occupations with relatively low unemployment rates are characterised by relatively high informality rates (or vice versa); consequently, increases in some occupations, for instance with low informality rates, might increase unemployment rates. Policymakers and training providers should be aware of this duality to provide adequate skills that genuinely improve people's employability.

9.2.3. Trends in the labour market

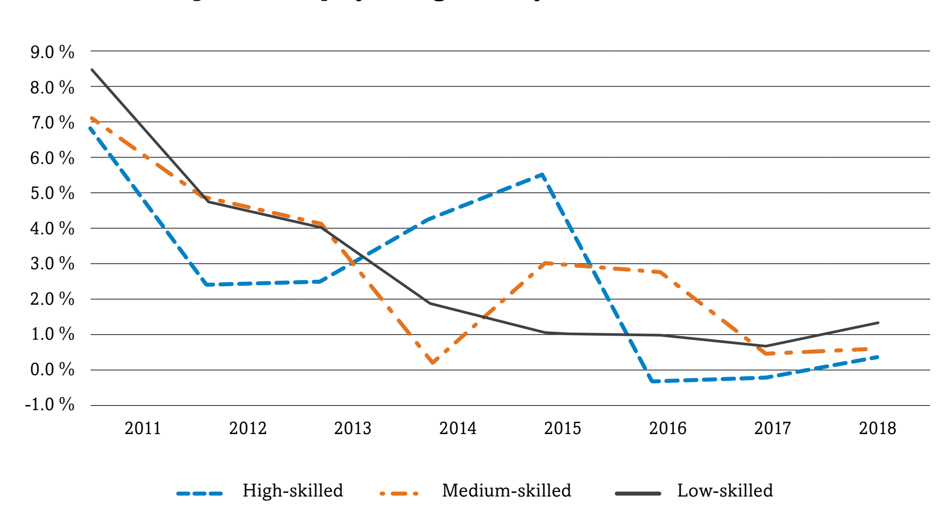
The above descriptive analysis shows the current state of the Colombian labour market. Nevertheless, it does not say anything about its dynamics. Given that possible changes might occur in the labour market, the conditions for a specific group of occupations might improve/worsen over time. Consequently, analysing labour market dynamics by occupations or skill levels will reveal whether there are favourable/unfavourable changes for a particular segment of the labour force. With this in mind, Figure 9.4 depicts the labour market composition of Colombian workers by skill level. As can be seen, the distribution of skills has remained approximately consistent over time (2010-2018). Low-skilled workers represent around 43% of the total of Colombian workers, followed by 32% medium-skilled and 25% high-skilled workers.

Figure 9.4. **Labour market composition of Colombian workers by skill level, 2010-2018**

Source: Author's calculations based on GEIH information, 2010-2018.

As shown in the figure, the overall structure of employed workers in Colombia has not considerably changed during this nine-year period. However, this composition has not changed because employment growth/decline has been relatively the same across all occupational groups. Figure 9.5 shows that, in general, employment growth for low-, medium-, and high-skilled occupations has decreased during the last decade. The decreasing trend of employment growth might be explained by labour supply and demand factors. It might be the case that the participation rate has declined or growth in demand has slowed during the last few years.

Figure 9.5. **Employment growth by skill level, 2011-2018**



Source: Author's calculations based on GEIH information, 2011-2018.

Chapter 3 has shown that the labour participation rate has been relatively consistent over the last decade (around 64%), while the unemployment rate has started to increase in the last four years. Figure 9.6 indicates how the unemployment rates for each skill level have increased. This evidence suggests that imbalances between labour supply and demand have been prevalent for all skill levels in the last few years.[135]

[135] Indeed, the Talent Shortage Survey released in 2019 by ManpowerGroup indicates that, in Colombia, there has been an increasing trend of talent shortages since 2011 (ManpowerGroup, n. d. 2019).

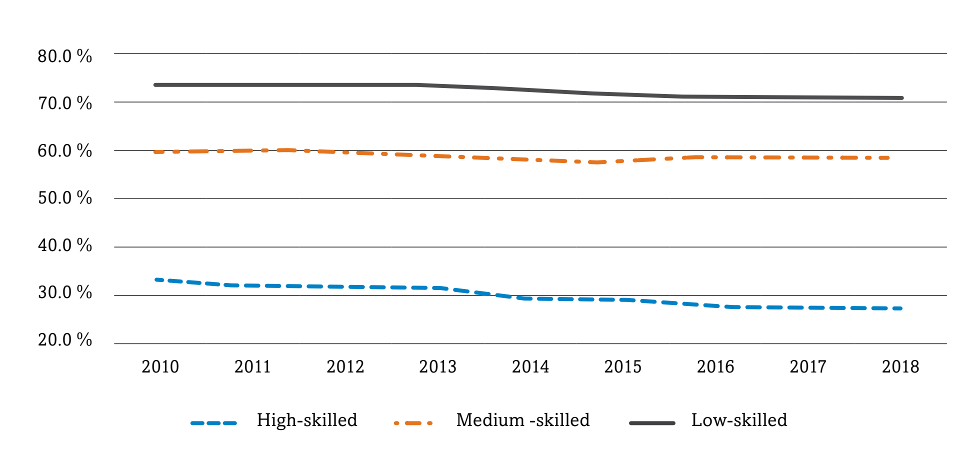
Figure 9.6. **Evolution of the unemployment rate by skill level, 2015-2018**

14.0 %
13.5%
13.0 %
12.5 %
12.0 %
11.5 %
11.0 %
10.5 %
10.0 %
2015
2016
2017
2018
High-skilled
Medium-skilled
Low-skilled

Source: Author's calculations based on GEIH information, 2015-2018.

Moreover, Chapter 3 has also demonstrated that informality rates have slightly decreased in the last four years. Figure 9.7 confirms in more detail how informality rates have slightly decreased for high- and low-skilled occupations, while for medium-skilled occupations this rate has remained relatively consistent over time. This result suggests that there has been an increase in skill oversupply, especially in low- and medium-skilled occupations over the last years.

Figure 9.7. **Evolution of the informality rate by skill level, 2010-2018**



Source: Author's calculations based on GEIH information, 2010-2018.

Thus, by considering the behaviour of unemployment and informality rates, in general, it can be observed that labour market outcomes have worsened across all skill groups since 2016. Specifically, the evidence indicates that low-skilled occupations show more signs of oversupply. Moreover, the recent increase in unemployment and informality rates (oversupply) suggests that there has been an increase in skill mismatch problems too.

However, a more detailed analysis might reveal that despite worsening employment conditions overall, for some occupations there have been improvements in terms of formal employment and real wages. For instance, a complete list of occupations in Appendix H (Table H.6) demonstrates that the employment growth trend has been positive between 2010 and 2018: around 47.4% correspond to high-skilled occupations, followed by 37.1% for medium-skilled and 15.5% for low-skilled occupations. Importantly, most of the occupations with the highest growth in labour demand (analysed in Chapter 6) are also in the mentioned list of occupations with positive employment growth (Table H.6). Such is the case, for instance, for "Computer network professionals," "Real estate agents and property managers," "Electronics engineering technicians," "Electronics engineers," and "Information and communications technology user support technicians." This evidence suggests that these occupations are in high demand.

Moreover, a complete list of occupations with a positive trend in real wages growth between 2010 and 2018 is also available in Appendix H (Table H.7). Around 42.6% of the occupations with a positive trend in wage growth are medium-skilled occupations, followed by 35.7% for high-skilled and 21.6% for low-skilled occupations. Most occupations with the highest growth in labour demand (as mentioned in Chapter 6) are found in the list of occupations with a positive trend in wage growth (Table H.7).

In summary, evidence suggests that Colombian workers face high rates of unemployment and informality, and, overall, their employment conditions have deteriorated since 2016. However, there are some segments in the labour market where formal employment and real wages have increased. This evidence might suggest that there are some occupations that are in high demand and might be at risk of skill shortages. Moreover, the considerable gap in the average wages of formal and informal workers by skill level indicates that informal workers and unemployed individuals have incentives to join the formal

economy. Potentially, occupations with skill shortages might be filled with the excess of supply from other occupations.

Nevertheless, further examination is required to determine whether there is a skill mismatch or not. For instance, the positive employment trend for some occupations might be due to improvements in labour market efficiency (e.g. reduction of search costs) rather than skill scarcity. Consequently, well-designed indicators of potential skill shortages are required to tackle labour market frictions, especially in Colombia, where skill mismatches (due to imperfect information) have been reported as one of the leading causes of relatively high unemployment and informality rates.

9.3. Measuring possible skill mismatches (macro-indicators)

Measuring skill mismatches is challenging. As pointed out by Bosworth (1993), "there is no one 'best way' to do it." Indicators that attempt to measure skill shortages might be affected by diverse factors; for instance, increased wages for a particular occupation might correspond to skill shortages or institutional and social factors (such as minimum wage increases or lower discrimination) (Shah and Burke 2003).

Consequently, the labour market literature has proposed different indicators to measure possible skill mismatches (see, for instance, European Commission 2015; MAC 2017; Mavromaras et al. 2013). The UK Migration Advisory Committee (MAC) has divided skill mismatch indicators in four categories (see Table 9.7 for a summary): employer-based, price-based, and volume-based indicators, as well as indicators of imbalance. As explained in Chapter 3, in Colombia, it is not possible to build employer-based macro-indicators because there are no sources of information (employer surveys) available. Instead, indicators of imbalance are used that refer to the vacancy to unemployment ratio (Beveridge curve). Briefly, the idea behind this indicator is that a high vacancy/unemployment ratio within an occupation or skill level might suggest that employers have difficulties in filling their vacancies, and vice versa.

Price-based indicators reveal that increases in real wages in a particular occupation are a possible sign of skill shortages. As explained in Chapter 2, in the basic labour market model, when there is an increase in labour demand

and labour supply is static, the real average wages tend to increase (given the relative labour shortage) to meet demand. Similarly, increases in employment and a reduction of the unemployment rate, etc. (volume-based indicators), are a sign of possible skill shortages.

Table 9.7. **Skill mismatch indicators**

Indicators set	Description
Employer-based indicators	Employer-based indicators are derived from surveys that ask employers direct questions about their demand for workers and their ability to recruit. Rising vacancy rates may suggest that employers are finding it hard to fill jobs. These data provide a valuable employer perspective; it is limited, however, since it only provides what employers choose to report.
Indicators of imbalance	Indicators of imbalance focus directly on vacancy levels within an occupation. A high vacancy/unemployment ratio within an occupation suggests that employers are having difficulty filling vacancies given the available supply of workers. Similarly, an increase in the average duration of vacancy indicates that employers are finding it more difficult to fill vacancies.
Price-based indicators	In case of a labour shortage, market pressure should increase wages, helping to raise supply and reduce demand, thus restoring labour market equilibrium. On this basis, rising wages within an occupation can be considered an indication of shortage.
Volume-based indicators	Increases in employment or increases in average hours worked may indicate rising demand and greater utilisation of the existing workforce, which could indicate shortage. Low or falling unemployment among people previously employed, or seeking to work, in an occupation may also indicate shortage (conversely, high unemployment amongst people seeking work in a particular occupation is an indicator that an occupation is not in shortage).

Source: MAC 2017.

As mentioned by MAC (2017), each set of indicators has advantages and disadvantages in measuring skill mismatches (see the following subsections). Consequently, both labour supply and labour demand information are necessary to determine where possible skill problems exist, and what labour demand requirements might not be fulfilled by the labour supply.

Nevertheless, in Colombia, a comparison between labour supply and labour demand information has been impossible because there has been no information about the labour demand or it has not been comparable with labour supply information, for example, not available at an occupational level (see Chapters 3 and 4). Therefore, one of the contributions of this document is that it makes

Colombian information about labour demand (job portals) and labour supply (household surveys) comparable in order to identify possible skill shortages.

In recent years, information from job portals has started to be considered as a source to measure possible skill shortages. For instance, the MAC has recently considered the use of this kind of information to design and update its skill shortage indicators. However, due to the collection of vacancy information provided by Burning Glass Technologies[136] (see Chapter 6), so far this source of information is only considered as a complement of the MAC indicators (MAC 2017). In contrast, Cedefop, which carries out the "Big Data analysis from online vacancies" project (see Chapter 4), has highlighted the potential of online vacancy information to provide information that reduces skill mismatches (Cedefop 2018). However, at the time of writing this book, the MAC's or Cedefop's skill mismatch indicators based on information from job portals have not been released.

Thus, Section 9.3 discusses how labour demand (job portals) and labour supply (household surveys) information can be used to determine possible skill shortages given the available sources of labour market information in a developing country such as Colombia.

9.3.1. Beveridge curve (indicators of imbalance)

The previous chapter has evidenced that information from job portals provides consistent information in terms of data representativeness with employment and unemployment series to reduce imperfect information issues in the labour market. Thus, it is possible to build indicators to continuously monitor and evaluate the match between labour supply and demand. Perhaps, one of the most well-known indicators for the evaluation of labour market matching is the Beveridge curve.

As mentioned in Chapter 3, the Beveridge curve relates vacancies to unemployment levels in order to determine how well, or inadequately, vacancies match unemployed workers. The curve is calculated by dividing the job openings

[136] Burning Glass Technologies count the number of advertised job postings as vacancies and do not consider (so far) the number of job placements a job advertisement might include (MAC 2017).

rate (the number of job placements as a per cent of total employment plus job placements) by the unemployment rate (total unemployed people divided by the total of employed and unemployed labour force):

$$\text{Beveridge curve} = \frac{\dfrac{\textit{job placements}}{\text{total employment + job placements}}}{\dfrac{\text{unemployed}}{\textit{labour force}}}$$

The points on the curve indicate the current business cycle of an economy.[137] Moreover, shifts to the right of the Beveridge curve indicate an increasing inefficiency of the labour market; in this scenario, there is a higher unemployment rate and a higher vacancy rate than before. This phenomenon is explained by an increase in labour market frictions, such as skill mismatches and labour mobility, among others. Shifts to the left of the Beveridge curve might indicate an increasing efficiency of the labour market; in this scenario, there are fewer frictions in the labour market allowing workers to match more easily a job vacancy (Bleakley and Fuhrer 1997). Theoretically, this curve slopes downward as the unemployment rate gets higher, the vacancy rate lower, and vice versa.[138]

Despite measuring labour market mismatch rather than skill mismatch, the Beveridge curve provides a first approach to assess the state of labour market matching. Moreover, it was expected that the Colombian Beveridge curve is strongly influenced by skill mismatches because the evidence found in Colombia, thus far (see Chapter 3), showed that skill mismatch problems are one of the most important causes of unemployment. Additionally, disaggregating the

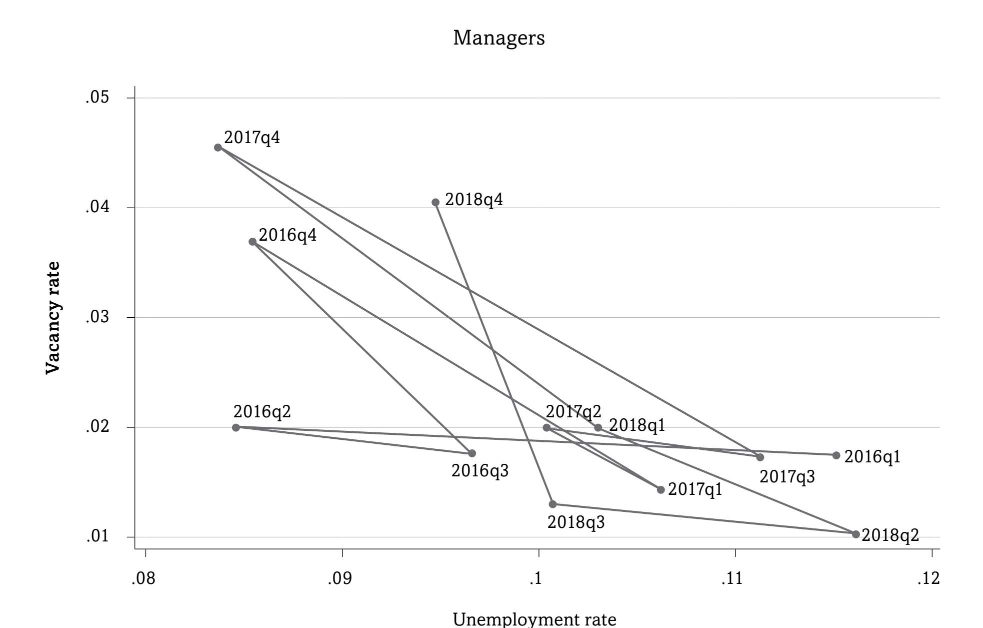
137 For instance, recession periods are characterised by a low vacancy rate and a relatively high unemployment rate (lower side of the 45° line), while periods of economic expansion are, generally, described by a high vacancy rate and a relatively low unemployment rate (upper side of the 45° line).

138 Empirically, different authors have demonstrated the downward slope of this curve in the US and other countries (Elsby, Michaels, and Ratner 2015). For Colombia, Álvarez and Hofstetter (2014) manually collected job advertisements in newspapers from 1976 to 2012 and estimated the aggregated Colombian Beveridge curve. They found, as expected, a downward slope between vacancy and unemployment rates. Consequently, the quarterly Beveridge curve calculated with the vacancy information for this book was also expected to show a downward slope.

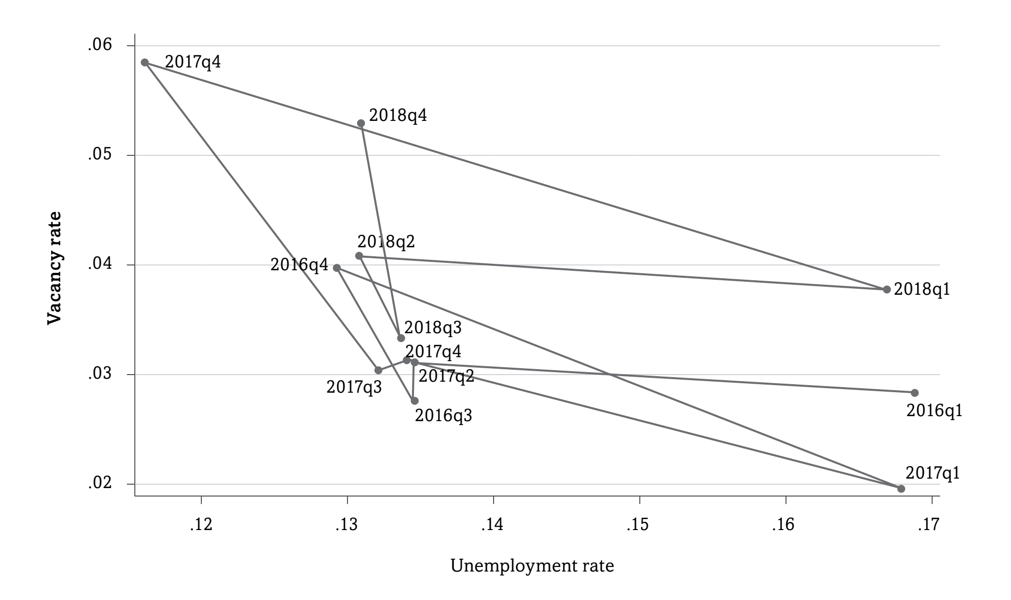
curve into occupational (one-digit) ISCO groups helped to determine which occupations might be experiencing more or fewer skill mismatch problems.

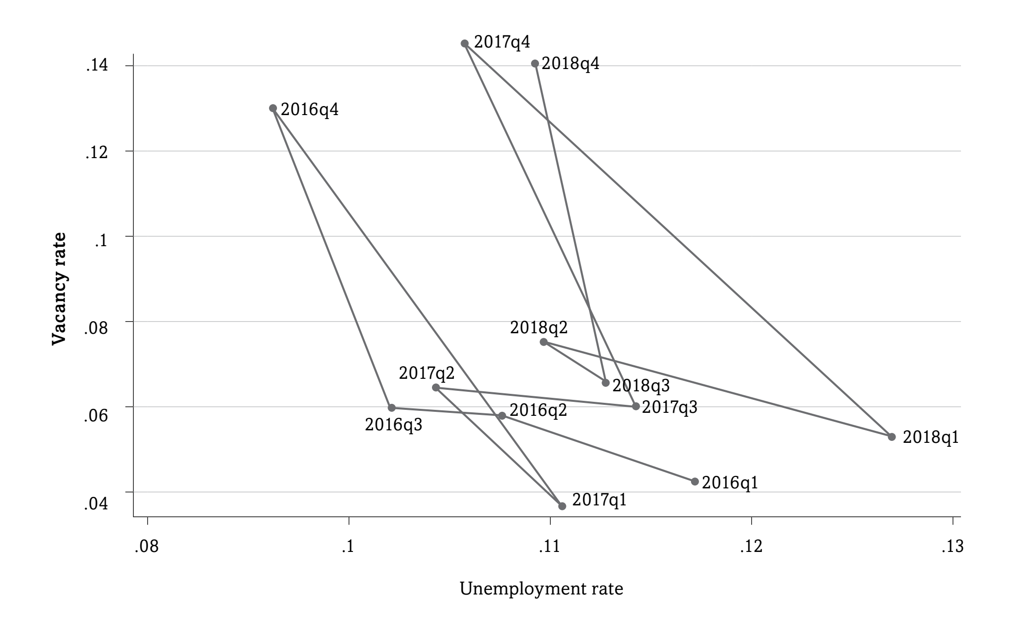
This document estimates the Beveridge curve at a one-digit occupational group level for the period 2016-2018 (similar to Turrell et al. 2018)[139] (see Figure 9.8). As can be observed here, the Beveridge curve is downward sloped by occupational groups; however, the occupational group "Skilled agricultural, forestry, and fishery workers" have some atypical points. This unexpected behaviour might be due to representativeness problems for the vacancy data within agricultural jobs (see Chapter 8). It is also worth considering that the GEIH information for this analysis does not take into account rural areas where most agricultural jobs are located.

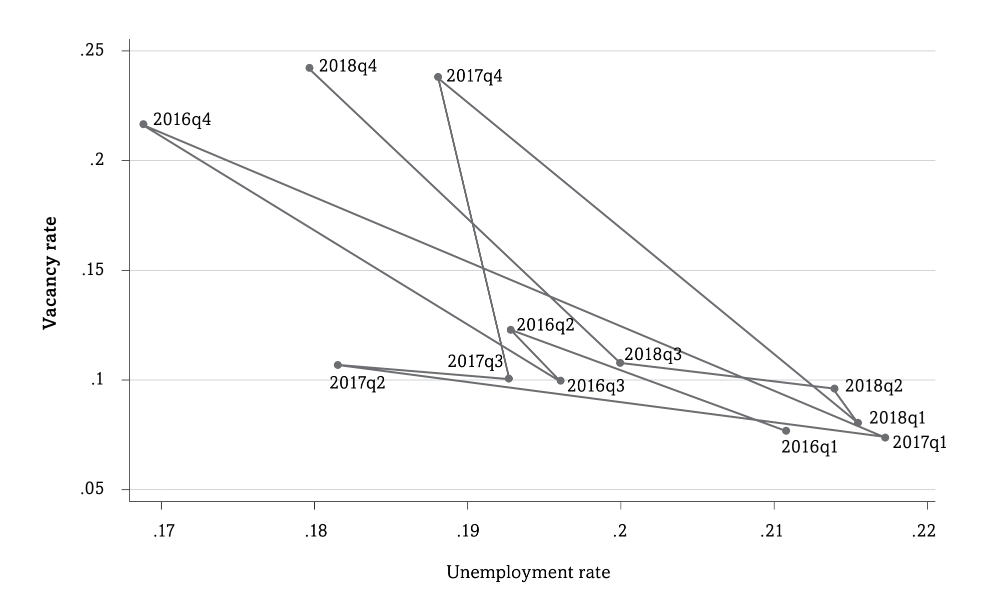
Figure 9.8. **Beveridge curve by (major) occupational groups**

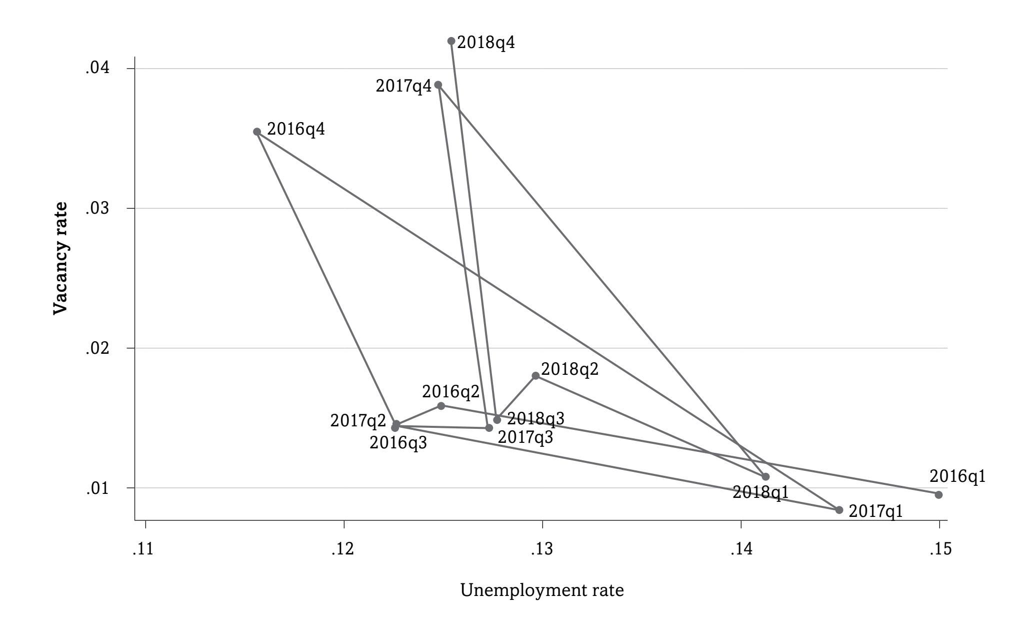


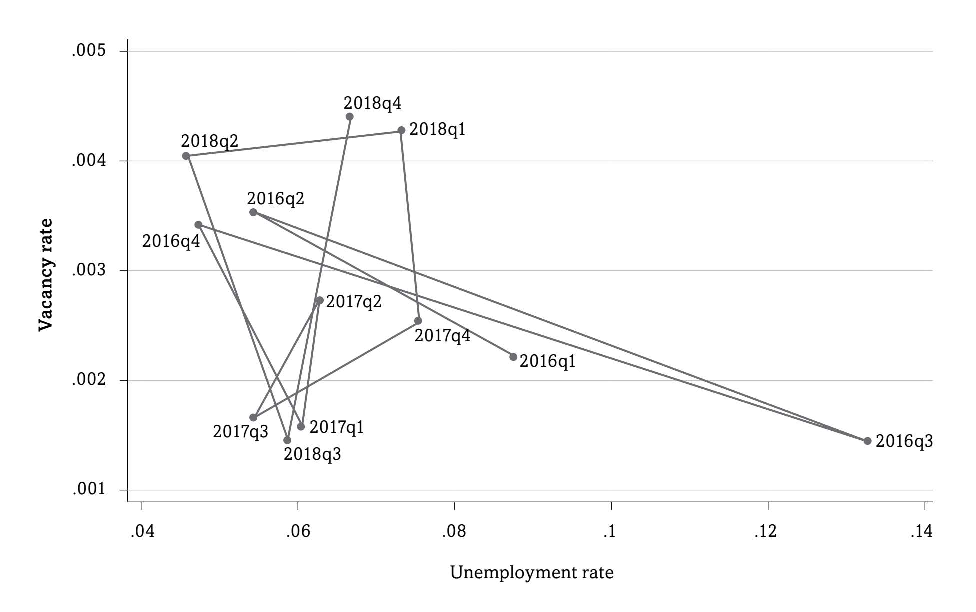
139 Given that the GEIH information has representativeness problems when data are excessively disaggregated (i.e. by quarter, four-digit occupational groups, etc. [see Chapter 3]), this book estimates the Beveridge curve at a one-digit occupational group level.

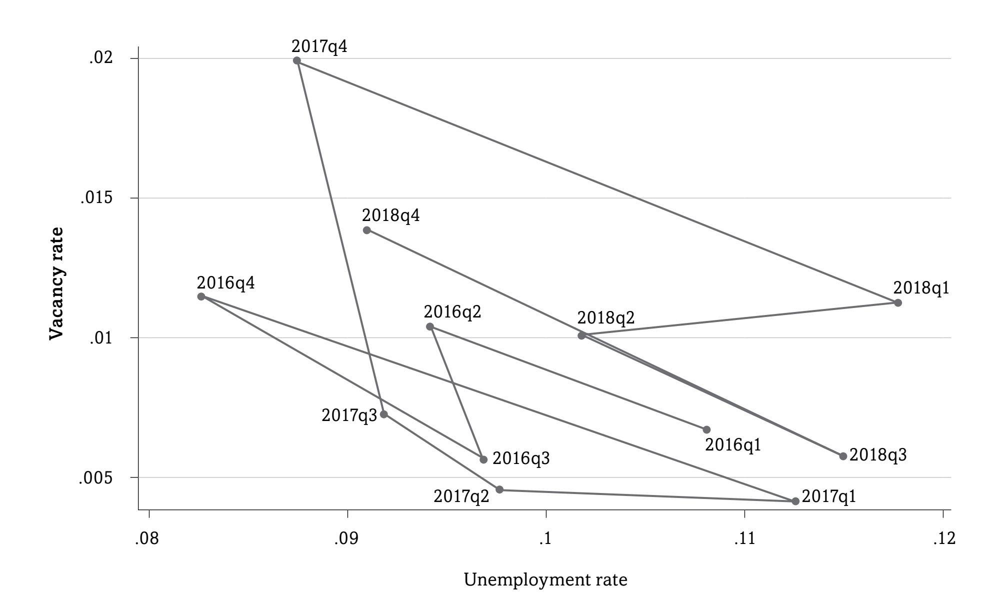

Professionals
Vacancy rate
.06
.05
.04
.03
.02
Unemployment rate
.12
.13
.14
.15
.16
.17
2017q4
2018q4
2018q2
2016q4
2018q1
2018q3
2017q4
2017q2
2017q3
2016q3
2016q1
2017q1

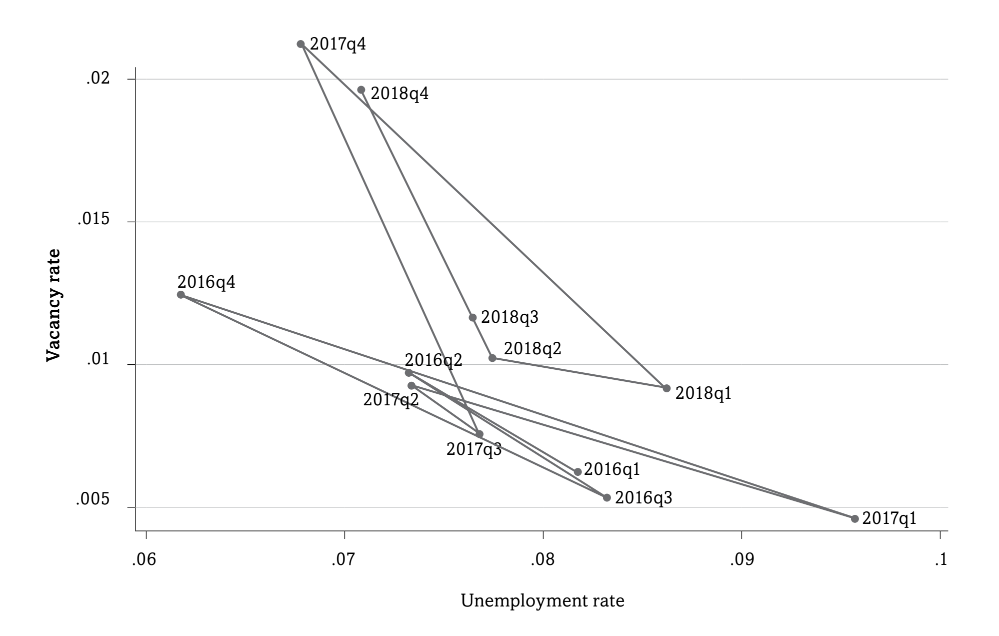

Technicians and associate professionals
Vacancy rate
.14
.12
.1
.08
.06
.04
Unemployment rate
.08
.1
.11
.12
.13
2017q4
2018q4
2016q4
2018q2
2017q2
2018q3
2017q3
2016q2
2016q3
2018q1
2016q1
2017q1

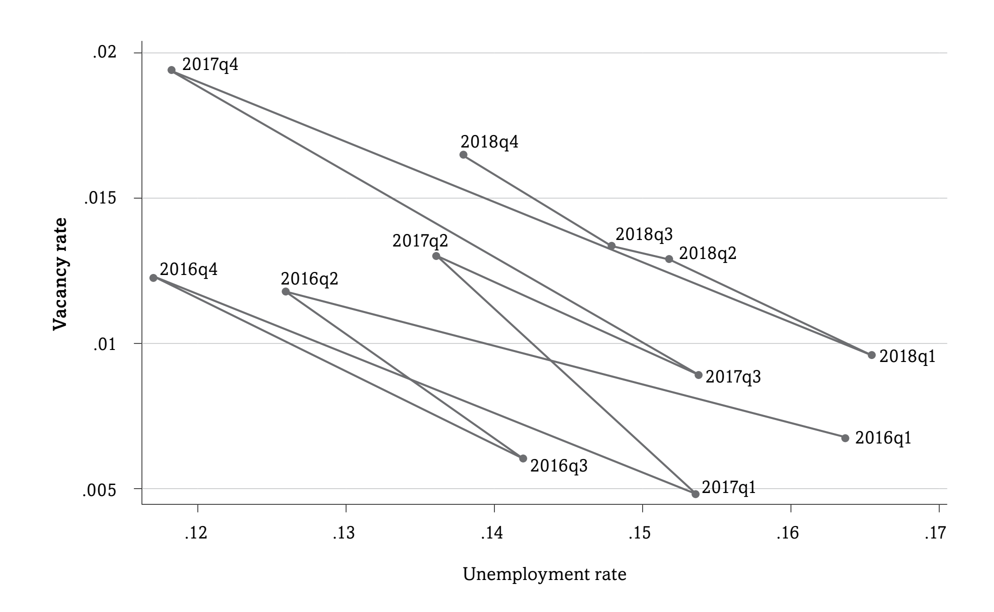

Clerical support workers
Vacancy rate
Unemployment rate
.25
.2
.15
.1
.05
.17
.18
.19
.2
.21
.22
2018q4
2017q4
2016q4
2016q2
2018q3
2017q3
2017q2
2016q3
2018q2
2018q1
2016q1
2017q1


Service and sales workers
Vacancy rate
Unemployment rate
.04
.03
.02
.01
.11
.12
.13
.14
.15
2018q4
2017q4
2016q4
2018q2
2016q2
2017q2
2018q3
2017q3
2016q3
2016q1
2018q1
2017q1


Skilled agricultural, forestry and fishery workers
.005
.004
.003
.002
.001
Vacancy rate
.04
.06
.08
.1
.12
.14
Unemployment rate
2018q4
2018q1
2018q2
2016q2
2016q4
2017q2
2017q4
2016q1
2016q3
2017q3
2017q1
2018q3


Craft and related trades workers
.02
.015
.01
.005
Vacancy rate
.08
.09
.1
.11
.12
Unemployment rate
2017q4
2018q4
2016q4
2018q1
2016q2
2018q2
2017q3
2016q1
2016q3
2018q3
2017q2
2017q1





Source: Author's calculations based on vacancy and GEIH information, 2016-2018.

In more detail, the Colombian Beveridge curve by occupational groups indicates two facts. First, the initial quarter of each year is characterised by higher unemployment rates and lower vacancy rates, while the last quarter of each year is characterised by lower unemployment rates and higher vacancy rates. This exercise shows that vacancies have, as expected, a positive relation with employment and a negative association with unemployment rates. Second, on average, the Beveridge curve for "Clerical support workers," "Professionals," and "Technicians and associate professionals" are farther from the origin (points [0,0] in Figure 9.8) compared to the other occupational groups. This evidence suggests that in these occupations there are likely to be higher labour market inefficiencies such as skill mismatches. Alternatively, the Beveridge curve for "Plant and machine operators and assemblers," "Craft and related trades workers," and "Managers" suggests fewer labour market frictions for those occupational groups.[140]

9.3.2. Volume-based indicators: Employment, unemployment, and vacancy growth

The Beveridge curve showed that occupational groups such as "Clerical support workers," "Professionals," and "Technicians and associate professionals" exhibit higher labour market frictions. However, the curve is affected by skill mismatches and other labour market issues (e.g. frictional unemployment, search costs, participation rates, etc.); consequently, further labour market indicators are needed to precisely determine possible skill shortages.

As previously shown in Table 9.7, volume-based and price-based indicators can be built to measure skill mismatches. For instance, the European Commission (2015) used the variation in employment and unemployment rates across skill levels as a measure of skill mismatch in the European Union. Increases or decreases in employment/unemployment rates are sought as a sign of skill shortages.

140 As mentioned above, at this moment, the vacancy data do not allow a long-term analysis of the Beveridge curve. So far, the present study helps to describe the current state of labour market frictions and compare them between occupational groups. However, in the future, when longer vacancy time series are available, it will be possible to calculate clearer shifts for the curve and, thus, observe increasing inefficiency/efficiency in Colombian labour matching.

This subsection focuses on volume-based indicators. As the name "volume-based" implies, these indicators are based on the number of working or unemployed people, or the number of hours worked.[141] Given the existing labour supply and new sources of labour demand information available in Colombia, it has become possible to estimate volume-based (and price-based) indicators of skill mismatch.

As mentioned in Chapter 3, one of the most developed approaches to measure skill mismatches can be found in the UK. Indeed, since 2008, the MAC has developed a conceptual framework and built 12 indicators of shortage using data for both labour demand and supply. Importantly, most of those indicators can now be adopted in Colombia given the updated information of labour demand and supply presented in this book. Thus, based on the system developed by the MAC and information available for Colombia, this document proposes the following volume-based measures to identify possible skill shortages.

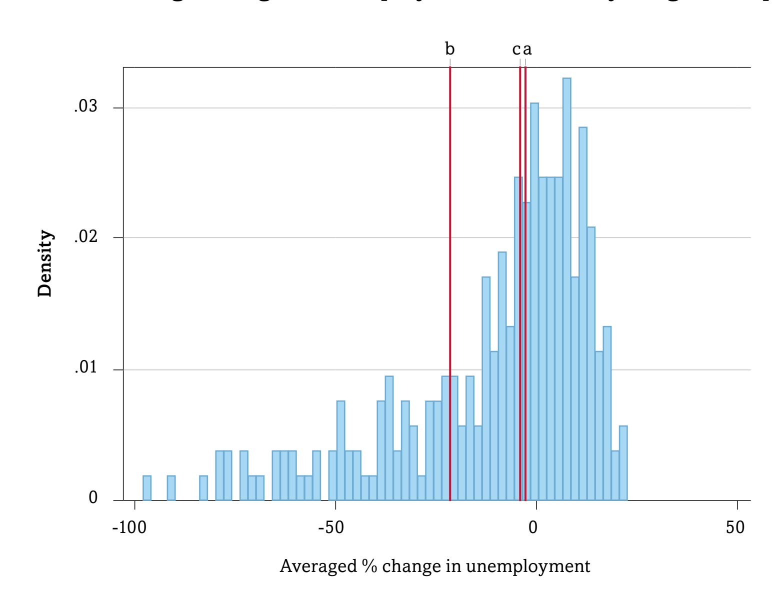
9.3.2.1. Percentage change in unemployment by sought occupation (three years)

As mentioned above, a decrease in the number of unemployed individuals is a sign that employers require relatively more people for a certain occupation, hence skill mismatch might arise. The GEIH provides information regarding sought occupations (job titles). However, given data representativeness issues, the annual percentage change in unemployment (and, in general, in most of the indicators that use household survey information) might excessively fluctuate and produce volatility in volume-based indicators, affecting thus the analysis of occupational changes. As proposed by the MAC (2017), one way to overcome this issue is by calculating skill shortage indicators averaged across three years. This three-year average identifies recent and less volatile occupational changes.

141 Increases in employment level or the average number of hours worked for an occupation might suggest a higher utilisation of a specific occupation and, hence, might indicate a potential skill mismatch. Conversely, a positive trend of unemployment might represent lower utilisation of a particular occupation; therefore, it might suggest that the occupation is not in shortage.

Figure 9.9 depicts the percentage change in unemployed individuals by sought occupation. Additionally, this and the following figures show the median, the third quartile, and the median plus 50 per cent of the median[142] (red lines a, b, and c, respectively). As will be discussed in Section 9.3.4, these thresholds help to determine at which point a specific indicator value should be considered as a sign of skill mismatch.[143] As observed in this figure, the median of this percentage change is -2.8%, and the third quartile is -21.4%. Moreover, the median plus 50 per cent of the median is -4.2%.

Figure 9.9. **Percentage change in unemployed individuals by sought occupation**



Source: Author's calculations based on GEIH information, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

142 The median is a measure of central tendency that is not affected by outliers.

143 The median and the third quartile are the most well-known measures of central tendency and dispersion. The median plus 50 per cent of the median is an alternative measure given that the median and the third quartile might be considered ambiguous or static thresholds to determinate skill mismatches (see Section 9.3.4). The median plus 50 per cent was selected (instead of, for instance, the median plus 10 or 90 per cent) to avoid this indicator from being similar to the median, or higher than the maximum value of a certain indicator. For instance, a particular variable can have the following values: 10, 30, and 50. The median of this variable is 30. The median plus 10 per cent (33) is similar to the median, while the median plus 90 per

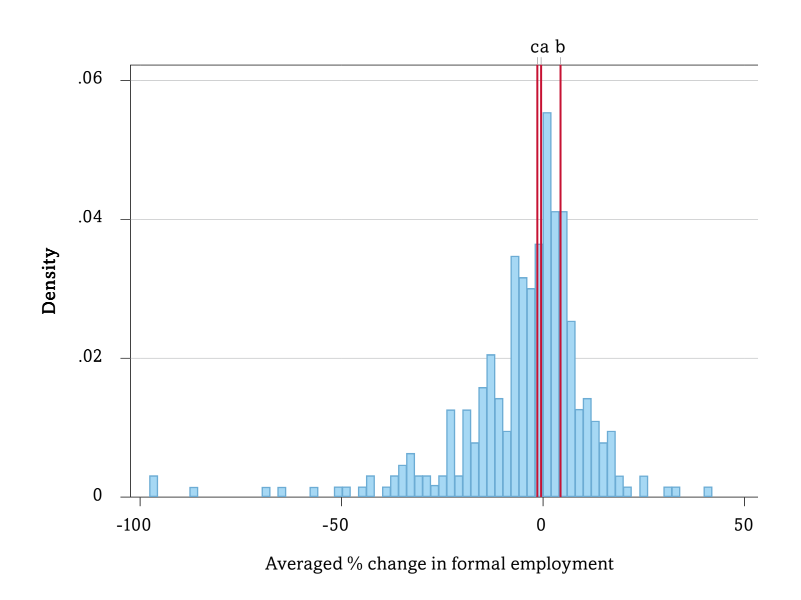
The distribution of this indicator shows that the number of unemployed individuals (by sought occupation) has increased for some occupations, while for other occupations it has decreased. This result suggests that employers have required relatively more people for certain occupations, while for other occupations labour demand shows signs of decline. However, the distribution is left-skewed and the mass of occupations is concentrated on the right of Figure 9.9, around a 0% change in unemployment. Moreover, the fact that the median is negative (-2.8%) indicates that more than half of the occupational groups experienced reductions in the number of unemployed individuals (by sought occupation). It is important to note that this result does not mean that the number of unemployed individuals (by sought occupation) has decreased over time. It might be the case that reductions in unemployment occurred in occupations with relatively few job seekers, while increases in unemployment happened in occupations with a relatively high number of job seekers.

9.3.2.2. Percentage change in formal employment (three years)

Contrary to the unemployment indicator, increases in the number of employees suggest that employers require relatively more people for a certain occupation and, hence, skill mismatch might arise. However, a distinction between formal and informal workers is required as growth in the level of employment might be due to people who could not find a formal job and opted for the informal economy instead. In this case, increases in the number of employees do not correspond to skill shortages (see Chapter 2). Instead, such increases would suggest that there is an oversupply for a specific occupation in the formal economy; consequently, given the proportion of informality in Colombia, it is important to calculate this indicator only for formal workers (see Figure 9.10).

cent is 55, which is higher than the maximum value of the variable. Instead, the median plus 50 per cent of the median is 45, and thus this threshold can be used to determine at which point a specific indicator value should be considered as a sign of skill mismatch.

Figure 9.10. **Percentage change in formal employment by occupation**



Source: Author's calculations based on GEIH information, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

As Figure 9.10 shows, the median of the percentage change in formal employment by occupation is -0.8%, the third quartile is 4.6%, and the median plus 50 per cent of the median is -1.3%. The percentage change in formal employment (controlling for some outliers) has a similar shape of a normal distribution curve centred at 0. This result indicates that a considerable proportion of occupations do not experience major changes in total formal employment numbers. However, certain occupations experience increases in the number of formal workers, suggesting that formal labour demand might have increased for particular segments of the labour market.

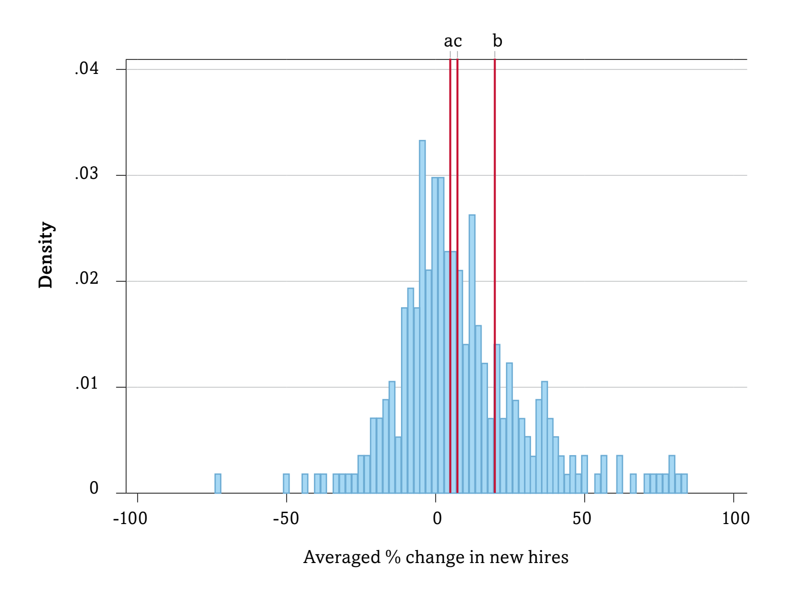
9.3.2.3. Percentage change in the proportion of formal workers in their job for less than a year: New hires (three years)

As discussed in the previous chapter, unemployment or employment levels might be influenced by different factors, such as lower dismissal rates or search costs, among others. The number of new hires, on the other hand, corresponds

to vacancies created by economic growth (net growth) and because people left their jobs (replacement demand). It is logical to think that when there is an increase in new hires, there is higher utilisation of the workforce for a specific occupation. Indeed, in Colombia, new hires have a strong correlation lag with the number of job openings (see Chapter 8). Consequently, new hires can be used as an indicator of possible skill shortages.

As in the case of the previous indicator, a distinction between formal and informal workers is required. Growth in the number of new hires might be due to people opting for the informal economy when they cannot find a formal job. Thus, this indicator is calculated by only accounting for the number of new hires in the formal economy (see Figure 9.11). As can be observed, the median, the third quartile, and the median plus 50 per cent of the median for this indicator is 4.6%, 19.6%, and 6.9%, respectively. The fact that the median is positive indicates that more than half of the occupational groups experienced increase in the number of new formal hires. Indeed, this distribution is slightly left-skewed.

Figure 9.11. **Percentage change in new hires by occupation**



Source: Author's calculations based on GEIH information, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

9.3.2.4. Percentage change in hours worked for formal employees (three years)

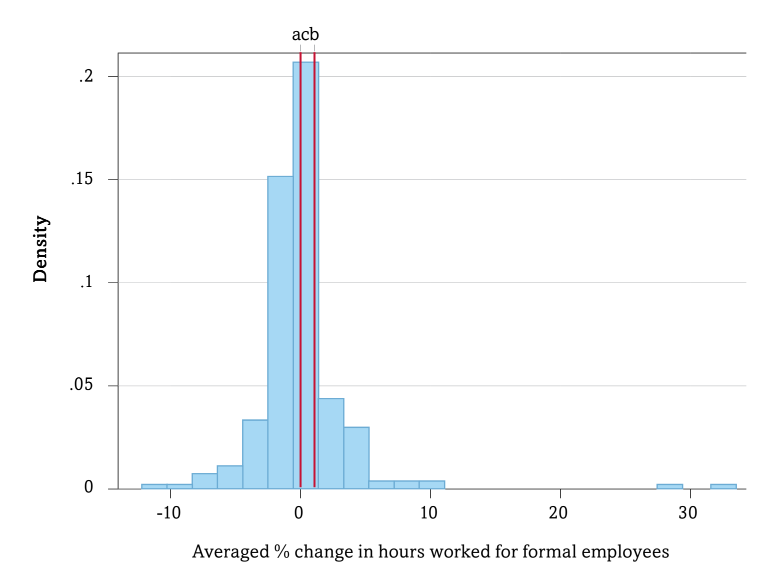
Alternatively, higher utilisation of the workforce for a particular occupation can occur through increases in the number of hours worked. It might be the case that employers do not find proper candidates to fill their vacancies; consequently, they might increase the number of hours worked by their current employees. Once again, a distinction between formal and informal workers is required: the number of hours worked in the informal economy might increase, while hours worked by formal workers might decrease. In this case, an increase in hours worked do not indicate that there is a possible skill mismatch.

Figure 9.12 illustrates the percentage change in hours worked for formal employees by occupation. The median of this indicator is around 0.00%, and the third quartile is 1.1%. Moreover, the median plus 50 per cent of the median is 0.01%. The percentage change in hours worked for formal employees (controlling for some outliers) has a similar shape to a normal distribution centred at 0. This result indicates that a considerable proportion of occupations do not experience major changes in hours worked. However, some occupations demonstrate increases in the number of hours worked, suggesting that formal labour demand might have increased for particular segments of the labour market.

9.3.2.5. Percentage change in job vacancy advertisements by occupation

As mentioned above, indicators based on labour supply might be influenced by other factors (e.g. labour participation) rather than a higher labour demand utilisation. Moreover, the previous chapters have shown that information from job portals represents occupational economic seasons and trends in Colombia’s labour demand; consequently, increases in the number of online job vacancy advertisements might be a sign of higher demand for a specific occupation and possible skill shortages. Thus, the annual percentage change in job vacancy advertisements might indicate a higher or lower use of the workforce by employers. Given that the vacancy information does not show high volatility in the period of analysis (2016-2018), the percentage change in job vacancy advertisements by occupation is not averaged across these three

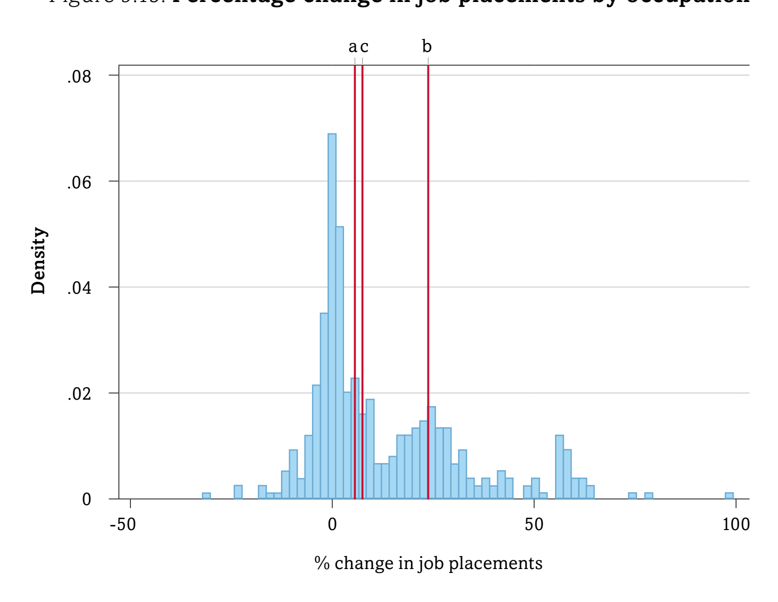
Figure 9.12. **Percentage change in hours worked for formal employees by occupation**



Source: Author's calculations based on GEIH information, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

Figure 9.13. **Percentage change in job placements by occupation**



Source: Author's calculations based on vacancy database, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

years. To some extent, how this vacancy information changes over one year guarantees that the use of a volume-based indicator is relevant in the short term for the identification of skill shortages.

As Figure 9.13 shows, the median of the percentage change in job placements by occupation is 4.1%, the third quartile is 23.4%, and the median plus 50 per cent of the median is 6.1%. In accordance with Chapter 7, percentage change in job placement distribution indicates that a considerable proportion of occupations do not experience major changes in labour demand (vacancies). However, the job placement distribution is right-skewed, which means that relatively few occupations experienced decrease in the number of advertised vacancies, while a higher number of occupations experienced an increase in job placements.

This subsection has discussed how proper volume-based skill mismatch indicators can be built using information sources available in Colombia. However, and in agreement with the MAC (2017) and Mavromaras et al. (2013), the identification of skill mismatches cannot be achieved by relying on just one indicator set. For instance, increases in the volume of employment or vacancies in specific occupations might be due to improvements in the search process (e.g. lower search cost) rather than real increase in the labour demand for a particular occupation. Thus, it is necessary to develop another set of indexes that use other labour market dimensions, such as prices, to complement volume-based indicators and indicators of imbalance.

9.3.3. Price-based indicators: Wages

As explained in Chapter 2, skill shortages might lead to increased wages. As labour demand increases for certain occupations, the current labour supply might not be enough to cover this higher demand; consequently, employers might have more difficulties in finding workers according to their requirements, and, hence, the wages of certain occupations might increase given the shortage of labour. Thus, information about wages might provide signs of skill shortages.

As in the case of volume-based indicators, in Colombia, the household survey (GEIH) provides information regarding the monthly and hourly wages of Colombian workers (prices), while information from job portals provides reliable information about vacancy wages (see Chapters 7 and 8). Therefore,

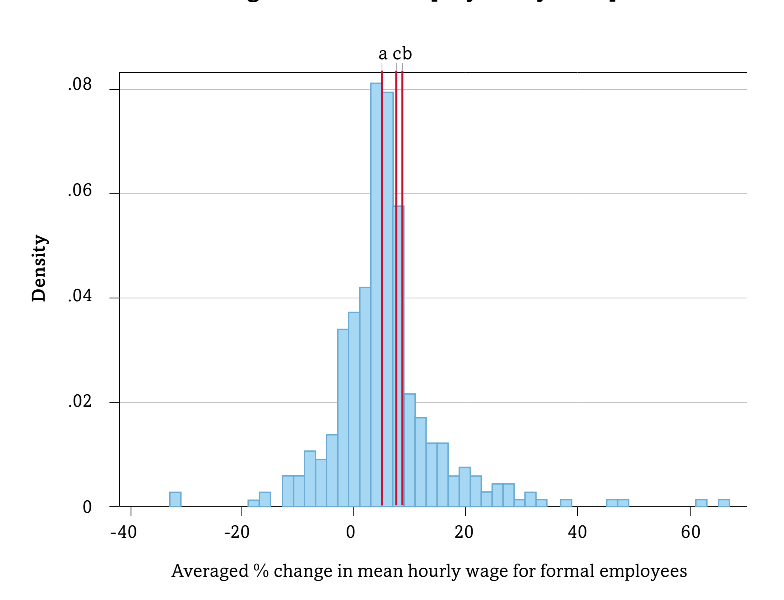
it is possible to build price indexes based on labour demand that are compatible with labour supply-based indicators, which might determine possible skill shortages.

9.3.3.1. Percentage change in median real hourly wage for formal employees (three years)

Estimating percentage change in wages might provide evidence regarding possible skill shortages. However, there are some points to consider in order to define this indicator in a way that it captures potential increases in labour demand. First, wage levels might increase over the years due to inflation. Second, the level of wages might be affected by the number of hours worked. Moreover, employers might react to skill shortages by increasing hourly salaries to improve a worker's productivity. Third, as discussed above, a distinction between formal and informal workers is required. Growth in wage levels might be due to increases in informal wages (the formal market might show an opposite trend), and, in this case, the percentage change in salaries might not necessarily suggest a skill shortage. Fourth, average wage figures might be affected by outliers. Finally, as is the case for volume-based indicators, household information might excessively fluctuate and produce volatility in price-based indicators, affecting the analysis of real wage changes at the occupational level.

To control for these issues, it is necessary to calculate the median value for the real wages (adjustment for inflation) of formal employees by dividing real salaries by the number of hours worked and averaging annual wage changes across three years. Figure 9.14 shows that the median, the third quartile, and the median plus 50 per cent of the median for this indicator are 4.9%, 8.4%, and 7.4%, respectively. The distribution of percentage change in mean real hourly wage for formal employees indicates that more than half of the occupational groups have experienced increases in real hourly wage. This result suggests that, for a considerable number of occupational groups, labour demand might have increased.

Figure 9.14. **Percentage change in mean real hourly wage for formal employees by occupation**



Source: Author's calculations based on GEIH information, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

9.3.3.2. Relative premium for an occupation: Controlling for education, region, and age

Alternatively, occupational shortages might indicate a relatively higher salary premium for those occupations compared with others. As mentioned above, companies tend to pay more to attract people with specific skills that are relatively scarce; therefore, the scarcer the supply in a particular occupation, the more likely a higher premium is offered for working in that occupation. Thus, the relative premium for an occupation can be expressed as follows:

$$\ln(w) = \beta_0 + \beta_1\, occupation_i + \varepsilon$$

Where w is wages, β_0 is the intercept, *occupation* is a dummy variable that takes the value of one when the premium is estimated for the occupation i, and ε is the error term.

However, the premium of a certain occupation compared to another might be affected by the characteristics of the geography or the people. For instance, the remuneration for an occupation might be affected by differences in the cost of living between regions—regions with a higher cost of life tend to pay higher wages, for example. Thus, it necessary to control for labour supply characteristics to estimate more precisely where occupational premium and skill shortages overlap. Nevertheless, there is a limit to the number of control variables because the higher the number of control variables, the more likely data representativeness issues will arise, given that household surveys might possess representativeness at a four-digit ISCO-08 level.

Thus, it is necessary to select the most relevant control variables to measure the relative premium for an occupation. One well-known approach to estimate a wage premium is the Mincer equation (see Chapter 2). This equation states that labour market income is a (linear and quadratic function) return on education and years of experience.

Usually, in the economic literature, the education variable is represented by years of education. This variable is available in the GEIH and can be used to estimate relative premium for an occupation. In contrast, the GEIH do not provide information regarding years of experience. However, a proxy frequently used for this variable is the worker's age. The older the worker, the more likely she/he will have more practical experience. Consequently, the worker's age is a correlated variable with the worker's experience. Moreover, as explained above, the level of prices in a region might affect the level of wages for a specific occupation; therefore, the region is an important variable to estimate relative premium for an occupation

Finally, high-skilled occupations tend to be better paid than low-skilled occupations (see Chapter 8); consequently, by definition, high-skilled occupations tend to have a higher premium and show signs of skill mismatch. Thus, to avoid comparisons between high- and low-skilled occupations, the relative premium was estimated by one-digit ISCO groups[144] (nine groups). Thus, the relative premium for an occupation can be expressed as follows:

$$\ln(w) = \beta_0 + \beta_1 occupation_{io} + + \beta_2 education_{io} + \beta_3 age_{io} + \beta_4 region_{io} + \varepsilon$$

[144] Higher levels of disaggregation can cause representativeness problems.

Where *w* indicates people's wages, β_0 is the intercept, and *occupation* is a dummy variable that takes the value of one when the premium is estimated for a person in the occupation and in the one-digit ISCO group *o*. The *education* and *age* variables are the worker's education (measured in years of education) and age, respectively, while *region* is the department[145] where the person works, and ε is the error term.

This equation controls for the most relevant elements while estimating salary premiums. Moreover, to estimate the relative premium of an occupation and to avoid representativeness issues and biases from informal economy, as much as possible, a pooled OLS (ordinary least squares) was conducted from 2016 to 2018 for formal Colombian workers.

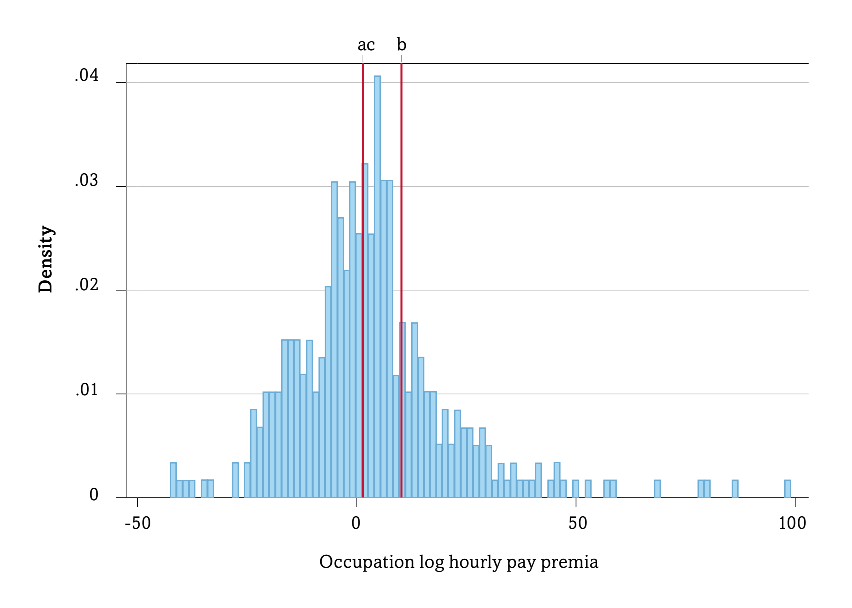
Figure 9.15 presents the distribution of regression coefficients. The median, the third quartile, and the median plus 50 per cent of the median occupation hourly pay premia are 1.9%, 10.8%, and 2.8%, respectively. There is a considerable number of occupational groups with positive hourly pay premia, which might indicate a shortage.

9.3.3.3. Relative vacancy premium for an occupation: Controlling for region and experience

As pointed out above, labour supply-based indicators might be influenced by other factors (e.g. labour participation) rather than a higher labour demand utilisation. Consequently, calculating the relative premium for an occupation using the vacancy database has an advantage because information comes from employer sources. Moreover, as showed in the previous chapter, the vacancy information is annually representative at a four-digit ISCO level for a considerable portion of occupations; thus, it is possible to annually estimate the relative vacancy premium for an occupation. To some extent, this estimation guarantees that the price-based indicator is relevant in the short term for the identification of skill shortages.

145 Amazonas, Antioquia, Arauca, Atlántico, Bogotá, Bolívar, Boyacá, Caldas, Caquetá, Casanare, Cauca, César, Chocó, Córdoba, Cundinamarca, Guainía, Guaviare, Huila, La Guajira, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, San Andrés and Providencia, Santander, Sucre, Tolima, Valle del Cauca, Vaupés, and Vichada.

Figure 9.15. **Occupational hourly pay premia**



Source: Author's calculations based on GEIH information, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

However, like any other indicator, the vacancy premium has limitations. Given the frequency of missing values, for instance, it is not possible (so far) to control for required years of experience. At most, it is possible to control whether a vacancy requires labour experience or not. Therefore, the relative vacancy premium for an occupation can be expressed as follows:

$$\ln(w) = \beta_0 + \beta_1 occupation_{io} + + \beta_2 diploma_{io} + \beta_3 experience_{io} + \beta_4 region_{io} + \varepsilon$$

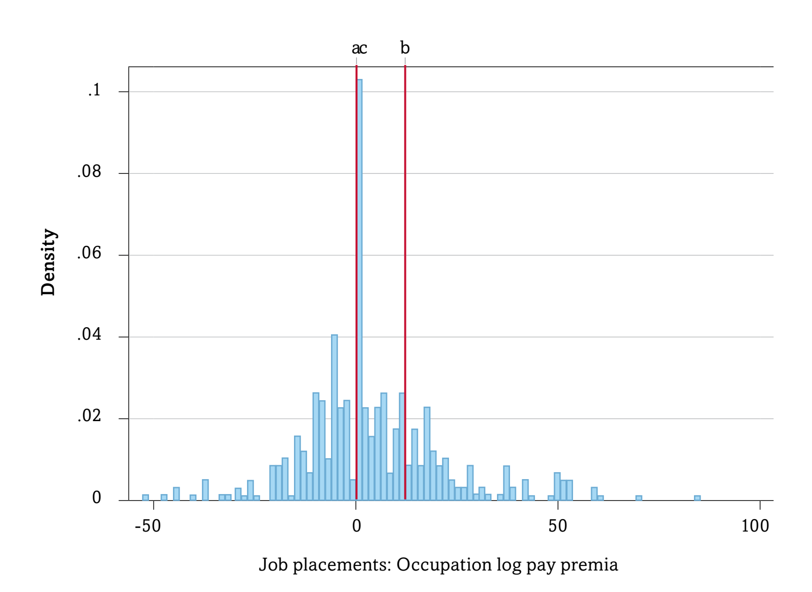
Where w is the vacancy's wages, β_0 is the intercept, and *occupation* is a dummy variable that takes the value of one when the premium is estimated for a vacancy in the occupation i and in the one-digit ISCO group o; *diploma* represents a set of dummy variables that indicate educational requirements (six categories, see Chapter 6, Table 6.2[146]). The variable *experience* is a dummy variable that takes the value of one if a vacancy requires experience and zero

[146] Due to frequency issues, the categories of specialisation, master's and doctoral degrees were grouped in one category: "Postgraduate."

otherwise, *region* is the department where the job vacancy is available, and ε is the error term.

Figure 9.16 presents the distribution of regression coefficients. The median, the third quartile, and the median plus 50 per cent of the median occupation pay premia within job placements is 0%, 12%, and 0%, respectively. As can be observed, the measures of central tendency tend to be positive in both Figure 9.15 and Figure 9.16. However, there are differences in the magnitude of the measures of central tendency for each one given that Figure 9.15 presents a higher hourly pay premium than Figure 9.16. As mentioned in Chapter 8, these differences might be explained by several reasons, such as the bargaining process or an employer's behaviour when posting wages in job advertisements. Despite their differences, Figures 9.15 and 9.16 show that there is a considerable number of occupational groups with positive hourly pay premia, which might indicate a skill shortage.

Figure 9.16. **Occupational pay premia within job placements**



Source: Author's calculations based on vacancy database, 2016-2018.

Note: Median (a), third quartile (b), and the median plus 50 per cent of the median (c).

9.3.4. Thresholds

Once the basic skill shortage indicators are established, the following step is to determine the threshold at which a specific index value should be considered as a sign of skill mismatch. In this regard, the MAC (2017) has taken part of an extended discussion regarding the adaptation of possible thresholds. As this institution has pointed out, there is no economic rule that fixes indicator thresholds. Consequently, given the MAC's recommendations and similarities between the MAC indicators and Colombian skill shortage indexes, this book considers the median, the quartile distribution, and median plus 50 per cent thresholds, which have been proposed by the MAC to determine at which value each indicator provides a sign of skill shortages.

The median and the quartile distribution are one of the most straightforward thresholds to determine at which value an indicator might suggest skill shortages. An occupation with values below or above the median might be considered as an occupation in deficit. However, independent of the economic cycle, quartiles (i.e. third quartile) and median thresholds will always provide the same number of occupations (i.e. 50% or 25% of the occupations) at risk of skill shortages (see MAC 2008). Consequently, even in situations where labour market works under perfect competition (see Chapter 2), these thresholds will always suggest occupational deficits.

Alternatively, the advantage of the median plus 50 per cent is that this threshold does not fix a specific number of occupations into skill shortage. Depending on the median value, the median plus 50 per cent threshold suggests a higher or lower number of occupations as being in short supply. However, this threshold might give inconsistent results. For instance, the median and the median plus 50 per cent of the percentage change in formal employment by occupation are -0.8% and -1.3%, respectively (see Figure 9.10). Occupations above these values could be considered at risk of skill shortages. Nevertheless, it is counterintuitive to conclude that those occupations with a negative value (between -1.3 and 0) in formal employment growth are at risk of skill shortages. Moreover, the median and the median plus 50 per cent might coincide when the median value of an indicator is at or closer to zero.

The fact that the median plus 50 per cent does not fix a certain number of occupations in short supply is an advantage that makes this indicator preferable

to others. However, in cases where the median plus 50 per cent threshold fails to provide consistent results, other rules will be considered alongside data to indicate possible skill shortages. Thus, the distribution of each indicator mentioned above needs to be analysed to select the most appropriate threshold. For the percentage change in unemployed individuals by sought occupation, the median plus 50 per cent is -4.2% (Figure 9.9). Decreases of more than -4.2% in unemployment by occupation suggest that employers require relatively more people for a specific occupation, hence skill mismatch might arise.

As mentioned above, the median plus 50 per cent of the percentage change in formal employment by occupation does not provide intuitive results because it suggests that occupations with negative formal employment values are experiencing skill shortages. Thus, in this case, when the third quartile value (4.6%) is selected to classify occupations, an increase of more than 4.6% in formal employment by occupation suggests shortages (Figure 9.10).

For the "new hires" indicator, the median plus 50 per cent provide intuitive results. Increases of more than 6.9% in formal hires by occupation suggest the occurrence of skill shortages (Figure 9.11). For the percentage change in hours worked for formal employees by occupation, the median plus 50 per cent gives the same value as the median (Figure 9.12). The median is almost zero, hence the median plus 50 per cent is close to zero. In such a case, the third quartile value (1.1%) is selected to classify occupations, and an increase of more than 1.1% of the hours worked of formal employees by occupation suggests skill mismatch.

The median plus 50 per cent threshold for the percentage change in job placements by occupation is 6.1% (Figure 9.13); therefore, increases in the percentage of online job vacancy advertisements of more than 6.1% are a sign of skill shortages. Likewise, the median plus 50 per cent threshold for the percentage change in mean hourly pay for formal employees by occupation is positive (Figure 9.14). Consequently, increases in percentage change of more than 7.4% regarding the mean hourly pay for formal employees suggest occupational deficits.

Regarding the occupational hourly pay premia of formal workers (Figure 9.15), the median plus 50 per cent threshold is 2.8%. Consequently, occupations with higher premia than 2.8% are potentially considered in short supply. In contrast, the median plus 50 per cent threshold for occupational pay premia

in job placements is the same as the median (Figure 9.16). Thus, in such cases, the third quartile value (12%) is selected to classify occupations and increases of more than 12.0% in the occupation pay premia for job placements suggest skill shortages. Table 9.8 summarises these indicators alongside their corresponding threshold values for an occupation to be considered in short supply.

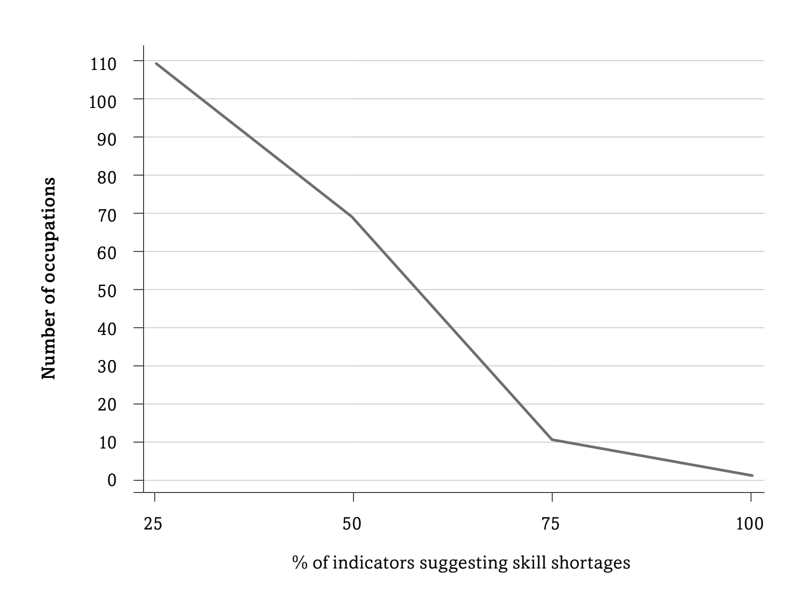
Table 9.8. **Skill shortage indicators and thresholds**

Indicator	Threshold type	Threshold value
% change in unemployed individuals by sought occupation	Median plus 50%	-4.2%
% change in formal employment	Top quartile (75)	4.6%
% change in the proportion of formal workers in their job for less than a year (new hires)	Median plus 50%	6.9%
% change in hours worked for formal employees	Top quartile (75)	1.1%
% change in job vacancy advertisements by occupation	Median plus 50%	6.1%
% change in median (real) hourly pay for formal employees	Median plus 50%	7.4%
Relative premium for an occupation, controlling for education, region, and age	Median plus 50%	2.8%
Relative vacancy premium for an occupation, controlling for region and experience	Top quartile (75)	12.0%

Source: Author's calculations based on vacancy database and GEIH, 2016-2018.

Once the measurement methods and thresholds have been established, the next step is to determine when an occupation shows strong signs of skill mismatch. As mentioned before, there is not an indicator that satisfactorily identifies every skill shortage. Instead, it would be excessively restrictive to expect that occupations in short supply will be identified by every indicator. Consequently, Figure 9.17 shows the number of occupations according to the percentage of indicators that suggest skill shortages. As can be observed, for instance, in 110 occupations, 25% or more of the indicators suggest skill shortages, while in 69 groups, at least, 50% of the indicators suggest skill mismatch issues.

Figure 9.17. **Number of occupations according to the percentage of indicators that suggest skill shortages**



Source: Author's calculations based on vacancy database and GEIH, 2016-2018.

Figure 9.17 helps to determine when an occupation shows strong signs of skill mismatch. As can be seen, for a relatively high number of occupations, half of the indices show signs of skill mismatch (69 categories). However, it is ambiguous to consider an occupation in skill mismatch when 50% of its indicators suggest skill shortages, while the remaining 50% do not. Moreover, the number of occupations with more than half of their indicators signalling skill shortages is considerably lower. This result indicates that thresholds above 50% might be adequate to distinguish skill mismatch occupations from other groups.

Nevertheless, in only 10 of the occupational categories, 75% or more indicators suggest skill shortages. Consequently, a threshold of 75% or more is excessively restrictive to classify occupations as exhibiting skill mismatch. Thus, to determine whether an occupation has shown enough evidence to be considered in short supply, this document suggests accepting a skill shortage if more than 50% of an occupation's indicators exhibit no missing values.[147] The MAC (2008) uses a similar condition to determine skill shortages in the UK.

[147] For some occupations, data were only available for a subset of indicators.

9.3.5. Skill shortages in the Colombian labour market

Table 9.9 lists the occupations that exhibit a strong sign of skill shortages. According to this table, 30 occupations are currently in short supply: 46.7% of the categories belong to high-skilled occupations, 36.6% to medium-skilled occupations, and 16.7% to low-skilled occupations. This evidence suggests that there exist formal labour market opportunities for people at all skill levels.

Table 9.9. **Occupations in skill mismatch**

Code	ISCO titles	Total indicators available	Total indicators passed	Percentage of indicators passed
2513	Web and multimedia developers	8	8	100.0%
2412	Financial and investment advisers	8	7	87.5%
2421	Management and organisation analysts	8	7	87.5%
2529	Database and network professionals not elsewhere classified	8	6	75.0%
7234	Bicycle and related repairers	8	6	75.0%
8154	Bleaching, dyeing, and fabric cleaning machine operators	8	6	75.0%
2521	Database designers and administrators	8	6	75.0%
7413	Electrical line installers and repairers	8	6	75.0%
2423	Personnel and careers professionals	8	6	75.0%
3118	Draughtspersons	8	6	75.0%
5113	Travel guides	7	5	71.4%
3432	Interior designers and decorators	7	5	71.4%
4313	Payroll clerks	7	5	71.4%
4221	Travel consultants and clerks	7	5	71.4%
4322	Production clerks	8	5	62.5%
5132	Bartenders	8	5	62.5%
4419	Clerical support workers not elsewhere classified	8	5	62.5%
2152	Electronics engineers	8	5	62.5%
8155	Fur and leather preparing machine operators	8	5	62.5%
5141	Hairdressers	8	5	62.5%

Code	ISCO titles	Total indicators available	Total indicators passed	Percentage of indicators passed
3259	Health associate professionals not elsewhere classified	8	5	62.5%
3141	Life science technicians (excluding medical)	8	5	62.5%
8321	Motorcycle drivers	8	5	62.5%
7314	Potters and related workers	8	5	62.5%
7214	Structural-metal preparers and erectors	8	5	62.5%
5312	Teachers' aides	8	5	62.5%
5112	Transport conductors	8	5	62.5%
2631	Economists	8	5	62.5%
2622	Librarians and related information professionals	8	5	62.5%
1342	Health services managers	7	4	57.1%

Source: Author's calculations based on vacancy and GEIH information, 2016-2018.

As can be seen, "Web and multimedia developers," "Financial and investment advisers," and "Management and organisation analysts" are occupations with the strongest signs of skill mismatch. It is important to note that occupations related to data, networks, and web professionals show clear shortage signs. These results confirm what has been said in Chapter 3, that labour market changes and new occupations might emerge; cases of occupations related to Big Data technologies (machine learning engineers, data sciences, Big Data engineers, among others) are representative examples.

The results from Table 9.9 strongly evidence that formal jobs have the best opportunities to absorb labour supply, which constitutes an important information for the Colombian government, education and training providers, and people in general in order to make policy decisions, provide training, and find employment. Consequently, based on the labour supply and demand model, in order to tackle informality and unemployment rates, it is necessary to make informal and unemployed people aware that jobs in certain occupations (see Table 9.9) offer the best chance to participate in the formal labour market, and to train people for these jobs. By considering people's characteristics and skill shortages, policymakers can design more precise public policies (e.g. routes of employment). For instance, given an informal or unemployed

person's occupation, it is possible to know which is/are the most similar job(s) to that person's current occupation where there is/are skill shortages. Based on this information, a person might decide to start applying for such jobs or (if necessary) to train and obtain the corresponding certification to apply for jobs experiencing skill shortages.

9.4. Detailed information about occupations and skill matching

The above section showed that by combining supply (GEIH) and labour demand (vacancy) information, it is possible to describe the structure and dynamics of the Colombian labour market and find convincing signs of skill mismatch issues. However, the advantage of online information from job portals is not limited to the provision of skill mismatch (macro) indicators. As shown in previous chapters, vacancy information provides detailed and updated information regarding employer requirements. Specifically, vacancy information provides detailed information about job requirements, and, hence, these data might function as a way to observe and reduce imperfect information regarding a country's skill needs. By monitoring relevant skills by occupations, the Colombian government and education and training providers might provide individuals with the proper skills demanded by employers. Moreover, people can make an informed decision regarding their career path. This section presents how detailed vacancy information might serve as a tool to improve labour market skill matching.

9.4.1. Skills

As demonstrated in Chapter 7, job descriptions for vacancies provide a rich source of information to analyse what skills are demanded by employers. However, it is important to clarify that employers do not provide a full list of skills needed for a specific occupation in each job vacancy description. First, providing a complete list of skills required for each vacancy would be a time-consuming task. Second, job descriptions tend to be concise and precise seeking to capture the attention of job applicants. Thus, employers only provide requirements that they consider to be the most essential ones for job

applicants in vacancy descriptions. Alternatively, employers might mention in the job description skills that are not easily found in job candidates. In both cases, the job vacancy description is a source that can be used to identify the most important skills in demand for a particular occupation, and the candidate who possesses those key skills will have better chances to obtain a job.

Consequently, a skills analysis might reveal the key skills and individual needs to apply for a certain job. Importantly, together with macro-indicators, job vacancy information can show which occupations are in short supply, as well as the key skills required to apply for those occupations. For instance, Table 9.10 presents five illustrative examples of occupations with skill shortages and what skills are frequently demanded for those occupations ("Web and multimedia developers," "Draughtspersons," "Travel guides," "Bicycle and related repairers", and "Productions clerks").[148]

Table 9.10. **Most demanded skills for occupations in skill mismatch**

ISCO title	Skill title	Skill type	Skill reusability level
Web and multimedia developers	SQL	Knowledge	Sector-specific
Web and multimedia developers	Database	Knowledge	Cross-sector
Web and multimedia developers	JavaScript	Knowledge	Sector-specific
Web and multimedia developers	Communication	Knowledge	Cross-sector
Web and multimedia developers	PHP	Knowledge	Sector-specific
Web and multimedia developers	Web programming	Knowledge	Sector-specific
Web and multimedia developers	MySQL	Knowledge	Sector-specific
Web and multimedia developers	Telecommunications engineering	Knowledge	Cross-sector
Web and multimedia developers	English	Knowledge	Transversal
Web and multimedia developers	Work in teams	Skill/competence	Transversal
Web and multimedia developers	Logic	Knowledge	Cross-sector
Web and multimedia developers	Visual Studio .NET	Knowledge	Sector-specific
Web and multimedia developers	LESS	Knowledge	Sector-specific
Web and multimedia developers	ASP.NET	Knowledge	Sector-specific
Web and multimedia developers	WordPress	Knowledge	Sector-specific

148 Given the quantity of the information and occupational categories, this subsection focuses on some illustrative cases.

ISCO title	Skill title	Skill type	Skill reusability level
Web and multimedia developers	Telecommunication industry	Knowledge	Cross-sector
Web and multimedia developers	Financial engineering	Knowledge	Cross-sector
Web and multimedia developers	Web analytics	Knowledge	Cross-sector
Web and multimedia developers	Sass	Knowledge	Sector-specific
Web and multimedia developers	Design process	Skill/competence	Cross-sector
Web and multimedia developers	Customer insight	Knowledge	Sector-specific
Web and multimedia developers	Spanish	Knowledge	Transversal
Web and multimedia developers	Drupal	Knowledge	Sector-specific
Web and multimedia developers	Solution deployment	Knowledge	Sector-specific
Web and multimedia developers	Control systems	Knowledge	Cross-sector
Web and multimedia developers	Computer programming	Knowledge	Transversal
Web and multimedia developers	Oracle WebLogic	Knowledge	Sector-specific
Web and multimedia developers	Business analysis	Knowledge	Cross-sector
Web and multimedia developers	ICT system integration	Knowledge	Sector-specific
Web and multimedia developers	Java (computer programming)	Knowledge	Sector-specific
Web and multimedia developers	Create an act	Skill/competence	Sector-specific
Web and multimedia developers	Business model	Knowledge	Occupation-specific
Web and multimedia developers	Data warehouse	Knowledge	Occupation-specific
Web and multimedia developers	E-learning	Knowledge	Sector-specific
Web and multimedia developers	DB2	Knowledge	Sector-specific
Web and multimedia developers	Office equipment	Knowledge	Sector-specific
Web and multimedia developers	Information architecture	Knowledge	Sector-specific
Web and multimedia developers	Maintain equipment	Skill/competence	Cross-sector
Web and multimedia developers	Design principles	Knowledge	Cross-sector
Web and multimedia developers	Xcode	Knowledge	Sector-specific
Web and multimedia developers	Analyse information processes	Skill/competence	Occupation-specific
Web and multimedia developers	Cisco	Knowledge	Sector-specific
Web and multimedia developers	Create model	Skill/competence	Occupation-specific

ISCO title	Skill title	Skill type	Skill reusability level
Web and multimedia developers	Create base for products	Skill/competence	Occupation-specific
Web and multimedia developers	Engineering principles	Knowledge	Cross-sector
Web and multimedia developers	Electrical engineering	Knowledge	Cross-sector
Web and multimedia developers	Office administration	Knowledge	Sector-specific
Web and multimedia developers	Object-oriented modelling	Knowledge	Sector-specific
Web and multimedia developers	Assess ICT knowledge	Skill/competence	Sector-specific
Web and multimedia developers	Search engines	Knowledge	Sector-specific
Web and multimedia developers	Innovation processes	Knowledge	Sector-specific
Web and multimedia developers	Microsoft Access	Knowledge	Sector-specific
Web and multimedia developers	Create solutions to problems	Skill/competence	Cross-sector
Web and multimedia developers	Systems Development Life Cycle	Knowledge	Cross-sector
Web and multimedia developers	Algorithms	Knowledge	Cross-sector
Web and multimedia developers	Information extraction	Knowledge	Sector-specific
Web and multimedia developers	Screen clients	Skill/competence	Cross-sector
Web and multimedia developers	Create software design	Skill/competence	Sector-specific
Web and multimedia developers	Perform business analysis	Skill/competence	Cross-sector
Web and multimedia developers	Electromechanics	Knowledge	Cross-sector
Web and multimedia developers	Data mining	Knowledge	Sector-specific
Web and multimedia developers	Financial statements	Knowledge	Cross-sector
Web and multimedia developers	Maintain database	Skill/competence	Cross-sector
Web and multimedia developers	Sales activities	Knowledge	Sector-specific
Web and multimedia developers	Assess customers	Skill/competence	Sector-specific
Web and multimedia developers	Portuguese	Knowledge	Transversal
Web and multimedia developers	ICT quality policy	Knowledge	Sector-specific
Web and multimedia developers	Information structure	Knowledge	Sector-specific
Web and multimedia developers	Write English	Skill/competence	Transversal

ISCO title	Skill title	Skill type	Skill reusability level
Web and multimedia developers	Perform data analysis	Skill/competence	Cross-sector
Web and multimedia developers	SQL Server Integration Services	Knowledge	Sector-specific
Web and multimedia developers	Apache Tomcat	Knowledge	Sector-specific
Web and multimedia developers	Perform system analysis	Skill/competence	Occupation-specific
Web and multimedia developers	Photography	Knowledge	Cross-sector
Web and multimedia developers	Show responsibility	Skill/competence	Cross-sector
Web and multimedia developers	Develop new products	Skill/competence	Sector-specific
Web and multimedia developers	Carry out sales analysis	Skill/competence	Sector-specific
Web and multimedia developers	Adobe Photoshop	Knowledge	Sector-specific
Web and multimedia developers	Lead a team	Skill/competence	Cross-sector
Web and multimedia developers	Assess object condition	Skill/competence	Sector-specific
Draughtspersons	Design drawings	Knowledge	Cross-sector
Draughtspersons	Communication	Knowledge	Cross-sector
Draughtspersons	Design process	Skill/competence	Cross-sector
Draughtspersons	Customer service	Knowledge	Sector-specific
Draughtspersons	Office equipment	Knowledge	Sector-specific
Draughtspersons	Customer insight	Knowledge	Sector-specific
Draughtspersons	Work in teams	Skill/competence	Transversal
Draughtspersons	English	Knowledge	Transversal
Draughtspersons	Trademarks	Knowledge	Cross-sector
Draughtspersons	Adobe Photoshop	Knowledge	Sector-specific
Draughtspersons	Information architecture	Knowledge	Sector-specific
Draughtspersons	Spanish	Knowledge	Transversal
Draughtspersons	Technical drawings	Knowledge	Cross-sector
Draughtspersons	Carpentry	Knowledge	Cross-sector
Draughtspersons	Give advice to others	Skill/competence	Transversal
Draughtspersons	Material mechanics	Knowledge	Cross-sector
Draughtspersons	Entertainment industry	Knowledge	Occupation-specific
Draughtspersons	Show responsibility	Skill/competence	Cross-sector

ISCO title	Skill title	Skill type	Skill reusability level
Draughtspersons	Geometry	Knowledge	Cross-sector
Draughtspersons	Innovation processes	Knowledge	Sector-specific
Draughtspersons	Adobe Illustrator	Knowledge	Sector-specific
Draughtspersons	Manage ICT projects	Skill/competence	Sector-specific
Draughtspersons	Lead a team	Skill/competence	Cross-sector
Draughtspersons	Monitor activities	Skill/competence	Cross-sector
Draughtspersons	Industrial software	Knowledge	Cross-sector
Draughtspersons	Instrumentation equipment	Knowledge	Cross-sector
Draughtspersons	Engineering principles	Knowledge	Cross-sector
Draughtspersons	Principles of mechanical engineering	Knowledge	Cross-sector
Draughtspersons	Design principles	Knowledge	Cross-sector
Draughtspersons	Algebra	Knowledge	Cross-sector
Draughtspersons	Maintenance and repair	Knowledge	Cross-sector
Draughtspersons	Manage personnel	Skill/competence	Cross-sector
Draughtspersons	Production processes	Knowledge	Cross-sector
Draughtspersons	Geographic information systems	Knowledge	Sector-specific
Draughtspersons	Digital printing	Knowledge	Sector-specific
Draughtspersons	Create model	Skill/competence	Occupation-specific
Draughtspersons	Create floor plan template	Skill/competence	Sector-specific
Draughtspersons	Publishing industry	Knowledge	Cross-sector
Draughtspersons	Food engineering	Knowledge	Sector-specific
Draughtspersons	Bridge engineering	Knowledge	Sector-specific
Draughtspersons	Visual Studio .NET	Knowledge	Sector-specific
Draughtspersons	Develop new products	Skill/competence	Sector-specific
Draughtspersons	Mathematics	Knowledge	Cross-sector
Draughtspersons	Design job analysis tools	Skill/competence	Occupation-specific
Draughtspersons	Information structure	Knowledge	Sector-specific

ISCO title	Skill title	Skill type	Skill reusability level
Travel guides	Customer service	Knowledge	Sector-specific
Travel guides	English	Knowledge	Transversal
Travel guides	Portuguese	Knowledge	Transversal
Bicycle and related repairers	Customer service	Knowledge	Sector-specific
Bicycle and related repairers	Maintenance and repair	Knowledge	Cross-sector
Production clerks	Work in teams	Skill/competence	Transversal
Production clerks	English	Knowledge	Transversal
Production clerks	Customer insight	Knowledge	Sector-specific
Production clerks	Textile industry	Knowledge	Cross-sector
Production clerks	Office equipment	Knowledge	Sector-specific
Production clerks	Customer service	Knowledge	Sector-specific
Production clerks	Characteristics of products	Knowledge	Sector-specific
Production clerks	Communication	Knowledge	Cross-sector
Production clerks	Production processes	Knowledge	Cross-sector
Production clerks	Medicines	Knowledge	Cross-sector
Production clerks	Maintain equipment	Skill/competence	Cross-sector
Production clerks	Pharmaceutical products	Knowledge	Sector-specific
Production clerks	Maintain machinery	Skill/competence	Cross-sector
Production clerks	Chemical products	Knowledge	Sector-specific
Production clerks	Construction products	Knowledge	Sector-specific
Production clerks	E-learning	Knowledge	Sector-specific
Production clerks	Mechanical tools	Knowledge	Cross-sector
Production clerks	Inspect quality of products	Skill/competence	Cross-sector
Production clerks	Maintenance and repair	Knowledge	Cross-sector
Production clerks	Footwear industry	Knowledge	Cross-sector
Production clerks	Machinery products	Knowledge	Sector-specific
Production clerks	Grade foods	Skill/competence	Occupation-specific
Production clerks	Trademarks	Knowledge	Cross-sector
Production clerks	ICT quality policy	Knowledge	Sector-specific

ISCO title	Skill title	Skill type	Skill reusability level
Production clerks	Perform business analysis	Skill/competence	Cross-sector
Production clerks	Flexography	Knowledge	Sector-specific
Production clerks	Data warehouse	Knowledge	Occupation-specific
Production clerks	Sales activities	Knowledge	Sector-specific
Production clerks	Give instructions to staff	Skill/competence	Cross-sector
Production clerks	Digital printing	Knowledge	Sector-specific
Production clerks	Exercise stewardship	Skill/competence	Cross-sector
Production clerks	Good manufacturing practices	Knowledge	Sector-specific
Production clerks	Dairy products	Knowledge	Sector-specific
Production clerks	Financial engineering	Knowledge	Cross-sector
Production clerks	Milk production process	Knowledge	Sector-specific
Production clerks	Mathematics	Knowledge	Cross-sector
Production clerks	Implement instructions	Skill/competence	Cross-sector
Production clerks	Carry out products preparation	Skill/competence	Sector-specific
Production clerks	Integrate ICT data	Skill/competence	Sector-specific
Production clerks	Design process	Skill/competence	Cross-sector
Production clerks	Identify customer requirements	Skill/competence	Cross-sector
Production clerks	Collect samples	Skill/competence	Sector-specific
Production clerks	Check the production schedule	Skill/competence	Sector-specific
Production clerks	ICT security standards	Knowledge	Sector-specific
Production clerks	Guarantee customer satisfaction	Skill/competence	Sector-specific
Production clerks	Perform system analysis	Skill/competence	Occupation-specific
Production clerks	Manipulate wood	Skill/competence	Cross-sector
Production clerks	Audit techniques	Knowledge	Cross-sector
Production clerks	Ensure information security	Skill/competence	Cross-sector

ISCO title	Skill title	Skill type	Skill reusability level
Production clerks	Animal food products	Knowledge	Sector-specific
Production clerks	Manage quality	Skill/competence	Transversal
Production clerks	Manage system security	Skill/competence	Sector-specific
Production clerks	Good laboratory practice	Knowledge	Cross-sector
Production clerks	Perform interviews	Skill/competence	Cross-sector
Production clerks	Operate video equipment	Skill/competence	Cross-sector
Production clerks	Liaise with government officials	Skill/competence	Cross-sector
Production clerks	Comply with schedule	Skill/competence	Cross-sector
Production clerks	Label foodstuffs	Skill/competence	Sector-specific
Production clerks	Compose condition reports	Skill/competence	Sector-specific
Production clerks	Weigh materials	Skill/competence	Sector-specific
Production clerks	Water pressure	Knowledge	Cross-sector
Production clerks	Database	Knowledge	Cross-sector
Production clerks	Present a cause	Skill/competence	Sector-specific
Production clerks	Order products	Skill/competence	Sector-specific
Production clerks	Upsell products	Skill/competence	Cross-sector
Production clerks	Develop new products	Skill/competence	Sector-specific

Source: Author's calculations based on vacancy and GEIH information, 2016-2018.

As observed in this table, the skill most demanded for "Web and multimedia developers" is SQL, followed by database (according to the ESCO skill definitions, database is "The classification of databases, that includes their purpose, characteristics, terminology, models, and use such as XML databases, document-oriented databases, and full text databases"), and JavaScript (the programming language of HTML and the web).

As mentioned in Chapter 2, technical skills are an important element for labour market matching. However, there are other types of skills (e.g. socio-emotional) that play a critical role in the matching process. With the information available from the vacancy data, it is possible to determine the most

frequently mentioned transversal skills. For instance, for "Web and multimedia developers," and "Draughtspersons," the most requested skills are for knowledge of English and for a person who can work in teams. Moreover, in some cases (such as "Production clerks") transversal skills such English, work in teams and communication are the most, or one of the most, frequently requested skills by employers.

Consequently, in general, the vacancy data provides important sector-specific, cross-specific, and transversal skills information. However, in some cases (e.g. "Travel guides" or "Bicycle and related repairers"), information from job portals provides a limited number of demanded skills. Due to the lack of observations, it is not possible to obtain a more comprehensive skill list for specific occupations.

With the information in Table 9.10, policymakers can design and induce training and educational programs that provide (at the very least) the most frequently demanded skills by employers. Likewise, with this information, education and training providers can update their curricula according to labour market needs. Furthermore, job seekers can make informed and better decisions in training and job search processes.

9.4.2. Skill trends

The results from Table 9.10 are essential to improve labour market skill matching. However, the utilisation of skills might vary over time. Especially, given rapid labour market changes (such as technological changes), some attributes might become more/less relevant than others to obtain a job. Increased demand of a particular skill for an occupation means that employers consider that characteristic more critical than others, or they are unable to find people with those requirements. Thus, analysing skill trends means identifying among the demanded skills the ones that are becoming more/less important for the labour market.

For illustrative purposes, Table 9.11 shows skills in demand with a positive trend for "Web and multimedia developers" from 2016 to 2018. Skills such as object-oriented modelling, create software design, Apache Tomcat, among others, exhibit a positive trend. Thus, special emphasis must be placed on providing those skills to "Web and multimedia developers." Moreover, the

results from Table 9.11 can be extended to other occupations. Consequently, the education and training system in Colombia—for example, career advisers, among others—can eventually improve the efficiency of addressing labour supply according to labour demand trends.

Table 9.11. **Skills with a positive trend for "Web and multimedia developers"**

Skill title	Skill type	Skill reusability level
Object-oriented modelling	Knowledge	Sector-specific
Create software design	Skill/competence	Sector-specific
Apache Tomcat	Knowledge	Sector-specific
Perform data analysis	Skill/competence	Cross-sector
Lead a team	Skill/competence	Cross-sector
Develop new products	Skill/competence	Sector-specific
Systems Development Life Cycle	Knowledge	Cross-sector
Perform system analysis	Skill/competence	Occupation-specific
Assess customers	Skill/competence	Sector-specific
ICT system integration	Knowledge	Sector-specific
Maintain database	Skill/competence	Cross-sector
ICT system integration	Knowledge	Sector-specific
Information extraction	Knowledge	Sector-specific

Source: Author's calculations based on vacancy and GEIH information, 2016-2018.

9.5. Conclusions

Unemployment and informality are widespread phenomena in the Colombian economy that affect people with different profiles. For instance, informality issues tend to be more prevalent in adults with a high school education (at most) who work in low-skilled occupations, while unemployment problem occurs with relatively more frequency in people younger than 29 years old who work in low- or high-skilled occupations. Furthermore, the considerable gap in the average wages of formal and informal workers by skill level indi-

cates that informal workers and those who are unemployed (regardless of their skill level) have incentives to join the formal economy. Thus, the Colombian labour market shows potential signals of skill mismatches at each skill level. However, low-skilled occupations tend to show more signs of oversupply: 1) a considerably higher informality rate compared to other skill groups; and 2) a high unemployment rate (slightly below the high-skilled unemployment rate). Consequently, skill shortages might be more frequent in medium- and high-skilled occupations.

Despite the high incidence of these phenomena, Colombia does not have a proper system (macro-indicators and skill monitoring) to reduce imperfect information issues by identifying possible skill shortages. Thus, this chapter has demonstrated that it is possible to develop a system for the identification of skill mismatches based on online vacancy information and household surveys in countries like Colombia.

Despite the relatively short period covered by the data, a Beveridge curve by occupational groups was estimated for Colombia from 2016 to 2018. This curve provides a macroeconomic context and indicates two facts: 1) the first quarter of the year for each occupation is characterised by higher unemployment and lower vacancy rates, while the last quarter of the year is characterised by lower unemployment and higher vacancy rates; and 2) on average, the labour market for "Clerical support workers," "Professionals," and "Technicians and associate professionals" has higher mismatches.

Moreover, the vacancy database, along with household surveys, can provide updated and precise indicators for the identification of skill shortages. However, it is important to note that "there is no one 'best way' to do it" (Bosworth 1993). Indeed, different approaches can be adapted from the literature (see Section 9.3). Given the relatively long experience of the MAC in designing skill mismatch indicators, as well as the vacancy and household survey information available for Colombia, this book concludes that the MAC indicators are a suitable framework for the Colombian context.

One of the most relevant elements for the adaptation of the MAC indicators to the Colombian context is the difference between the formal and informal economy in Colombia. Increases in the level of employment might be due to increases in the number of informal workers. In this scenario, growth in the number of employees does not correspond to skill shortages. On the contrary,

this outcome indicates that oversupply exists for a particular occupation. Thus, given the size of the informal economy in Colombia, skill indicators should be estimated by only considering formal workers.

The skill mismatch indicators for Colombia demonstrate that 30 occupations are currently in short supply. This list is composed of high-skilled occupations (46.7%), followed by medium- (36.6%) and low-skilled occupations (16.7%). Therefore, the evidence suggests that there exist formal labour market opportunities for people with different profiles in terms of age, education, and work experience, amongst others.

These results have a high relevancy for Colombia because they allow a continuous and consistent monitoring of skill shortages at a relatively low cost and over a short time period. However, the scope of job vacancy information is not limited to the estimation and improvement of skill mismatch indicators at an occupational level: one of the greatest advantages of using job portal data for a system of skill mismatch identification is that these sources enable the analysis of skills demanded over time for a certain occupation. For instance, for "Web and multimedia developers," there is an increasing demand for the skills of object-oriented modelling, create software design, and Apache Tomcat, among other skills.

Based on these results, 1) policymakers and education and training providers can quickly promote and update policies and curricula, according to the current occupational labour demand structure and specific skills required; 2) the government and career advisers, among other related professionals, can design better routes to employment based on people's profiles and employer requirements; and 3) job seekers can receive relevant information regarding occupation shortages and, more importantly, the corresponding skills in demand. In this way, unemployed and informal people can make better and informed decisions about their training and job search processes.

In summary, vacancy information is a valuable resource that provides consistent and unique (unmet) labour demand data for a considerable set of non-agricultural, non-governmental, non-military, and non-self-employed ("business owners") occupations in the urban and formal economy. With a systematic analysis of this information, economic agents can reduce unemployment and informality rates by taking informed and better decisions according to up-to-date labour market needs.



10. Conclusions and Implications

10.1. Introduction

This book investigated to what extent a web-based model of skill mismatches could be developed for Colombia, where information regarding the labour market is relatively scarce. In particular, this research sought to answer how and to what extent novel sources of information from job portals might be used to inform policy recommendations, especially to address two major labour market problems in Colombia, which are relatively high unemployment and informality rates (9.4% and 47.2% in 2017, respectively). Indeed, the Colombian economy had the second highest unemployment rate in the Latin American region in 2015, and the informality rate was around 1.4 percentage points higher than the Latin American average in the same year (ILO 2016b) (see Chapter 3).

Under the model of perfect competition, the over- or undersupply of skills (skill mismatches expressed in informality and unemployment) only arise over the short term (Bosworth, Dawkins, and Stromback, 1996). Consequently, this mode cannot explain the high and persistent informal and unemployment rates that exist in countries like Colombia. The conditions needed for perfect competition seldom exist because agents usually possess imperfect information about offered skills and those in demand (Garibaldi 2006; Reich, Gordon, and Edwards 1973; Stiglitz et al. 2013). This failure in the labour market can create skill shortages. Workers with the skills demanded by employers are more likely to be engaged in the formal economy, while workers without the "proper" skills have more chances of being absorbed by the informal economy or being unemployed. Consequently, an economic model with imperfect information better explains labour market outcomes in countries like Colombia (Chapter 2).

The evidence gathered for this document suggests that one of the leading causes of high unemployment and informality rates in Colombia is due to skill shortages. Employers, labour market experts, and national and international institutions agree that, in general, Colombian job seekers do not have

appropriate skills to fill available vacancies (OECD 2015a; ManpowerGroup, n.d.; Arango and Hamann 2013). People make decisions about human capital investments and look for jobs based on imperfect information; hence, they do not meet employer requirements. Moreover, employers do not have perfect information about the skills possessed by potential workers and where they can be found (Desjardins and Rubenson 2011; Oyer and Schaefer 2010) (Chapter 2). Despite the high incidence of these phenomena, Colombia does not have a proper labour market analysis system to identify possible skill shortages and current employer requirements regarding skills. One main reason for the absence of a skill mismatch identification system is that the collection of labour demand information through traditional sources (employer surveys) is costly, in terms of resources and time. Additionally, the available data do not provide detailed characteristics of the labour demand for certain occupations or skills over time, and thus it is not possible to draw comparisons between labour demand and supply information (Chapter 3).

Recently, the use of online job portals as a source of information has attracted the attention of researchers and policymakers (Kureková, Beblavy, and Thum, 2014), since job portals seem to provide quick and relatively low-priced access to analyse labour demand information. Importantly, this kind of data might be more relevant in contexts where employers experience difficulties in filling job vacancies; in these instances, information from job portals might be the only data available to analyse labour demand for skills in order to address labour supply limitations according to employer requirements. Information from job portals has considerable potential to fill information gaps in different countries, in real time and at a low cost (Chapter 4).

Like any other source of information, Big Data sources need to be examined to avoid biases and determine the scope of these data. For instance, given internet penetration rates and the type of job portal users who have access to online vacancies, online vacancy data might have some representativeness limitations. The existence of those potential limitations, though, does not necessarily invalidate the use of job portals for labour market insights. In fact, an identification of the biases and limitations of online vacancy data helps us to understand the topics on which the analysis of online vacancy data might have a higher and decisive impact, such as the design of labour market public policies and academic research.

However, in general, little has been done to investigate, in depth, the advantages, limitations, and possible uses of job portals to tackle skill mismatches in different countries. Thus, this study contributes to a better understanding of the advantages and limitations that job portals can provide towards addressing public policy issues or academic research problems by: 1) showing how skill mismatches help to generate unemployment and informality in Colombia; 2) providing a better understanding of these phenomena through the development of a framework that uses matching theory and Big Data concepts (Chapters 2 to 4); 3) developing methods and proposing criteria to collect and organise information from job portals in a consistent and efficient way (Chapters 5 to 6); 4) testing the internal and external validity of online vacancy data for economic analysis (Chapter 8); 5) providing a detailed and novel analysis of unmet Colombian labour demand (Chapters 7 to 9); and 6) determining skill shortages based on job portals and household surveys using an updated occupational classification (household survey occupational information has been updated to ISCO-08 thanks to the methodologies conducted in this book) (Chapter 9).

This book has used a variety of quantitative methods to collect, clean, organise, categorise, compare, and analyse a large amount of labour demand vacancies and labour supply information. Specifically, web-scraping methods were implemented to systematically and automatically collect vacancy information from three main Colombian job portals over three years. Once this information was collected, text mining techniques were used to structure the vacancy database (Chapter 5). As skills and occupational data are one of the most important features of this vacancy information (and also one of the most difficult features to organise for statistical analysis), a combination of machine learning as well as automated and manual classification methods were conducted to obtain consistent occupation and skills variables. Once the vacancy information had been structured, the next step was to impute the values of essential variables for statistical analysis, such as education and salaries (Chapter 6).

The following careful steps were taken to test the vacancy database. First, a statistical analysis provided a description of the vacancy database. This analysis also served as an initial approach to comprehend the structure and the dynamics of the Colombian unmet labour demand in detail (Chapter 7). However, one of the main concerns of using information from job portals for statistical purposes is that little is known about the possible biases and limits

of these sources of information (Chapter 4). Thus, internal and external validity tests were conducted to determine these biases and limitations (Chapter 8). Once data limitations were established, quantitative methods focused on providing a labour market analysis were proposed using a combination of demand and supply information to monitor skill shortages and address labour supply according to employer requirements (Chapter 9).

This chapter presents the conceptual contributions in Section 10.2, while the main contributions of this work to methodology, such as the collection and validation processes to evaluate the vacancy database, are discussed in Section 10.3. The empirical contributions (e.g. main findings of skill mismatches and skill requirements) are presented in Section 10.4. The implications of this study for policymakers, education and training providers, and job seekers are described in Section 10.5. Limitations and further research are discussed in Sections 10.6 and 10.7, respectively. Finally, Section 10.8 presents the concluding statement to this research.

10.2. Conceptual contributions

As Kureková, Beblavy, and Thum (2014) state (Chapter 4), the debate regarding the use of online sources (e.g. job portals) for labour market analysis is flawed, one reason being that any source of information has certain limitations or biases (census, surveys, the internet, etc.). Additionally, most studies that use vacancy data obtain information from private companies, and their methods (and corresponding changes over time) for collecting such data remain in a "black box" (see, for instance, Lima and Bakhshi 2018, or Turrell et al. 2019). Consequently, these studies are not able to explain in detail how data were obtained and processed, nor the challenges and limitations of consolidating an unmet labour demand database. Moreover, given that these sources of information are relatively new and there are challenges to test the validity of these data, there is a lack of debate concerning the types of research questions information from job portals can provide consistent and valuable data for so as to conduct an adequate labour market analysis (see Chapters 4 and 8).

Thus, this book offers an important contribution to the debate about whether data from job portals can be used more extensively, and to what extent they

provide reliable results. Specifically, this research determines whether online vacancy information can provide key and reliable information to manage labour supply according to labour market requirements. Despite different concerns about the use of information from job portals for labour market analysis, this study found that, with the proper techniques, it is possible to obtain online vacancy information of a relatively high quality (Chapters 5 to 9).

As discussed earlier in Chapter 4, the quality framework and guidelines provided by the OECD establish seven dimensions in order to evaluate the quality of data in a specific database: relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence (OECD 2011, pp. 7-10) (Chapter 4). Table 10.1 presents the findings of this research regarding this framework.

Table 10.1. **OECD quality framework and vacancy data**

Criteria	Result
Relevance	The online vacancy database is (at the very least) a relevant source to gather information about skill mismatches and job requirements in the Colombian labour market (Chapters 7 to 9).
Accuracy	The vacancy database broadly describes the structure of urban unmet labour demand, except for self-employed (“business owners”), informal, governmental, military, and agricultural occupations (Chapters 7 to 9).
Credibility	This book shows evidence that it is possible to consolidate a consistent vacancy data in accordance with proper statistical standards (Chapters 4 to 6).
Timeliness	This criterion is one of the most important advantages of information from job portals compared with other sources of information. Once the algorithms and statistical procedures are established to collect information from job portals, it is possible to analyse employer requirements almost immediately after the information is created. This document has shown that vacancy information helps to guarantee that skill shortage indicators are relevant in the short term (Chapter 9).
Accessibility	This book demonstrated that a consistent vacancy database can be consolidated from job portals and, potentially, this information and derived results can be made accessible to the public (Chapters 5 to 9).
Interpretability	Given the theoretical framework, definitions, target population, and representativeness of this study, the interpretability of the vacancy database considerably improves. This book has shown that the analysis of vacancy and household survey information can be combined to produce consistent and easy-to-interpret indicators for skill shortages (Chapter 9).
Coherence	The vacancy data provides internally and externally consistent results. For instance, information from job portals adequately represents the “real” trends and economic seasons of the total number of job placements in Colombia (Chapter 8).

These criteria have evidenced that this document addresses the key issues of vacancy data. Consequently, the research demonstrates that the concept and sources of Big Data (in this case, from job portal sources) can provide consistent results to orient public policies (e.g. identifying skill shortages) (see Chapters 7 to 9). Importantly, it also shows that, using the proper techniques, online vacancy data can fulfil conceptual requirements to be considered as high-quality data for labour market analysis (see Chapters 4 and 10).

Moreover, this book makes a conceptual contribution by showing, first, how skill mismatches help to create informality and unemployment, and, second, by providing a better understanding of these problems through the development of a framework that uses matching theory and Big Data concepts (e.g. concepts of informality, unemployment, skills, imperfect information, the causes of labour market segmentation, web scraping, etc.) (see Chapters 2 to 4). As will be discussed in Section 10.3, by considering some important elements, such as the size of the informal economy, this document has defined and estimated skill mismatch indicators according to the Colombian context. The specific contributions of the book are now explained.

10.3. Contributions to methodology

As mentioned above, most research that uses vacancy information obtains data from private companies, and their methodologies, challenges, and changes for consolidating a vacancy database remain in a "black box." Consequently, these studies are not able to discuss and overcome important concerns such as data quality, representativeness, internet penetration rates, etc., of the online sources used, as happens in the case of job portals (see Chapter 4). There does not exist a discussion about a comprehensive methodology for collecting, consolidating, and analysing job portals in order to tackle skill mismatch issues. This lack of discussion concerning methodology has undermined the credibility of a consistent and useful potential source of labour demand information from job portals.

This book makes an original contribution by developing an extensive and novel mixed-methods approach to process and analyse the advantages and limitations of information from job portals; more specifically, to address

labour supply according to employer requirements. Next, the main aspects of this contribution to methodology are described in more detail.

Vacancy information is available from multiple websites. However, collecting job advertisements from each job portal in the country might not be an optimal approach to build a vacancy database. First, each job portal has its own HTML structure. Consequently, it is necessary to develop and update an algorithm for each website to extract labour demand information. Including every job portal in the country is inconvenient due to limited resources (time, money, human and computational capabilities, etc.). Second, employers can advertise the same vacancy in one or more job portals. Consequently, the larger the number of websites scraped in order to consolidate the vacancy database, the more chances are that duplication problems arise.

Conversely, to consider just one job portal is also problematic because one website might be focused on a specific part of the labour market, hence results from that source might not be representative of the economy. Additionally, some job portals might provide false or low-quality vacancy information (see Chapter 5). Therefore, not every job website is good enough to provide vacancy information and, hence, it is critical to establish rigorous criteria to select the job portals that can provide a less biased understanding of labour demand. Consequently, Chapter 5 proposes three criteria to select the most relevant job portals to better capture the dynamics of the labour market: 1) volume (number of advertisements available), 2) website quality (structure and number of variables), and 3) traffic ranking (number of users). Based on these criteria, a vacancy data base was built for Colombia.

So far, web scraping techniques are the best way to obtain labour vacancy information from job portals. However, there are challenges to consider regarding this technique. First, conducting web scraping requires a deep understanding of programming (such as R and Python, among others) and a knowledge of the architecture of each job portal selected in the sample (HTML). Second, each website has a unique HTML structure and, as a consequence, different algorithms are required to be programmed that automatically and periodically collect information from each website. Third, websites might change over time, thus, algorithms need to be updated whenever there is a change in the HTML structure of the sample websites. Fourth, given such changes, the number of job portals in the selected sample might also vary over time. Thus, to overcome

the above issues, this document programmed different algorithms that automatically and periodically collect information from each selected job portal. These algorithms were periodically revised to ensure their proper functioning.

This book discussed and applied different methods to consolidate a consistent vacancy database for economic analysis and public policy advice. One key strength of this mixed-methods approach is that it overcomes linguistic (such as gendered words in Spanish) and orthographic (misspelled words) issues, and merges different datasets that have the same identification keys (e.g. company names in the vacancy and Business Registry database) in order to compile a homogenous vacancy database for analysis (Chapter 5). By using these methods, it was possible to organise (homologate) information from different job portals into a single database for statistical analysis.

Importantly, this research considerably contributes to current understandings by applying this novel mixed-methods approach to identify skills and occupations in online job announcements, which would otherwise be complex to collect via other means. First, in countries like Colombia, information regarding skills is widespread in online job advertisements, and employers do not use pre-defined categories to describe required skills. Moreover, in this country, a national official skill dictionary is unavailable to identify which words in vacancy descriptions correspond to a specific skill. This document proposed the use of international dictionaries such as the ESCO (a multilingual classification of European skills, competencies, qualifications, and occupations) to build a methodology that identifies the skills being demanded in each job advertisement.

With the implementation of text mining techniques (such as stop words, stemming, etc.), each pattern in the skills dictionary was searched in each job vacancy advertisement. A skill variable took the value of one if a certain pattern in the skills dictionary was found in the advertisement, and zero otherwise. Consequently, it was possible to identify the skills required by Colombian employers via job portals. Additionally, this book found that with the help of similar text mining techniques as those mentioned above, it is possible to identify country-specific or new skills that are not listed in the ESCO dictionary but are mentioned in online job vacancy descriptions. By doing so, this research provides an innovative and comprehensive methodology to automatically categorise skills from job announcements (Chapter 6).

Second, job titles are the backbone of this vacancy analysis. The categorisation of job titles into occupations is one of the most critical procedures because this occupational variable summarises the main characteristics of labour demand. Literature has developed different methods and algorithms to classify job titles into occupations, such as manual coding, classifiers, machine learning algorithms, etc. (Jones and Elias 2004; Gweon et al. 2017). Although machine learning algorithms have recently attracted the attention of researchers across disciplines (e.g. economics and statistics), these automatic methods do not classify, so far, the entire job title sample. In some countries like Colombia, a training database (data to train occupational classification algorithms) is not yet available, although it is a fundamental input to conduct machine learning techniques. This book recommends, as a first step, the combination of manual, semiautomatic, and automatic classification techniques to accurately classify as many job titles as possible. Furthermore, this document proposes an extension of a machine learning algorithm (nearest neighbour algorithm) that takes into account not only available job titles, but also skill requirements to increase the accuracy level and the number of coded job titles (Chapter 6).

Another critical issue concerns duplication. As vacancy data are collected from different websites, some job advertisements can appear on more than one job board or even on the same job portal. This study has argued that an n-gram-based approach (which is not sensitive to minor changes in string variables) is the best method to minimise duplication issues (Chapter 6). This document has shown that once the variables are organised and categorised, it is possible to impute values. Indeed, it has been demonstrated that differences between imputed and non-imputed wages are minimal (Chapters 7 and 8).

Thus, this novel mixed-methods approach has improved data collection and aided a better understanding of the methodological changes required to obtain information from job portals. As a product of this robust methodology, it has been possible to test the validity of an online vacancy database to analyse possible skill shortages in a developing country such as Colombia.

Like any other source of information, the vacancy database has its limitations. This book has addressed one of the most critical concerns regarding job portal data, which is the internal and external validity of these sources of information (i.e. their internal and external consistency or data representativeness). To test internal validity, this research proposed to compare different but

correlated variables: wage and vacancy distribution by educational, experience, and skill groups. This comparison enabled the understanding of possible biases or identifying errors in data collection in the most relevant variables for a skill mismatch analysis (Chapter 8).

Ideally, to examine the external validity of vacancy information collected from job portals, it is required to have an updated census of vacancies, which details the total vacancies available in a given country. Nonetheless, carrying out and maintaining an updated census is expensive in terms of time and money. Indeed, countries with less restricted budgets, such as the UK, also face issues to collect a vast amount of vacancy information. As mentioned by the UK ONS, "It is not feasible to survey every business in the UK" (ONS 2019). Thus, in Colombia, there is no available census of vacancies or a similar database (see Chapter 3). Therefore, testing the representativeness of the vacancy database in Colombia is challenging because it is not possible to utilise a vacancy census or any official information to comprehend the total number of vacancies (statistical universe).

Despite these various difficulties, this book has provided a methodology to carry out the external evaluation of the vacancy database with sources of information available in the country. A "traditional" (aggregated and static) comparison of occupational structures was conducted between the vacancy database (demand) and total employment from the GEIH (supply). However, this exercise was limited, given that total employment is composed of the number of job matches, while information from job portals is the total of the net and replacement labour demand (see Chapter 8). Thus, this document has developed further validity tests as part of its methodological contribution.

First, a static comparison was proposed between the distribution of wages in the vacancy database and the GEIH household survey. Second, given that vacancy information was collected for three years, a time series comparison was conducted between the number of vacancies and employed and unemployed people as well as new hires in order to determine whether economic seasons could be observed in the vacancy database. These comparisons evaluated whether the vacancy data are representative of the labour market structure, and whether these sources of labour demand information reflect the economic seasons and trends of the Colombian labour market (Chapter 8). As a result of these validity tests, it was concluded that information from job portals reflects

Colombian economic seasons and trends for a considerable set of occupations (see Section 10.4).

Moreover, this book has made an important methodological contribution to the measurement and analysis of skill mismatches. First, it has proved that (with proper techniques and corresponding precautions) job portals, along with official sources of information (such as household surveys), can be used to provide an overview of the labour market, while skill shortage indicators enable the monitoring of skill requirements over time. Indeed, indicators used (for instance) by the MAC can be improved with high-quality vacancy information; this helps to have labour market insights over a relatively short period, and such information can fill information gaps (Chapter 9).

Second, this document contributes to the ongoing debate on skill mismatch measurements by taking informality into account. As mentioned in Chapter 2, in Latin America, especially in Colombia, informality rates are relatively high, and skill mismatches are a vital explanation of these results. A considerable part of employment growth might be due to people who were not able to find a formal job and opted for the informal economy. In this case, increases in the number of employees do not correspond to skill shortages; this information suggests that there is an oversupply in the formal economy. Therefore, skill mismatch indicators need to control for informality to avoid misleading results (Chapter 9).

In summary, this book has provided a comprehensive guide to collect and analyse vacancy information from job portals, and to test its validity. This framework is particularly useful for countries like Colombia where testing and comparing the representativeness of a vacancy database based on online sources is more challenging because labour demand information collected by traditional methods such as vacancy surveys is, at best, relatively scarce.

10.4. Empirical contributions

As outlined in Chapter 4, little attention has been paid to possible research questions that information from job portals can help to answer for different countries, even considering the particular limitations and biases of these sources. In Colombia, given various problems related to collecting detailed and

representative labour demand information through surveys, the occupational structure of labour demand, its dynamics, and employer skill requirements are relatively unknown. Due to this lack of labour demand information and the use of outdated occupational classifications in household surveys, it had previously not been possible to conduct a combined analysis of labour demand and supply and estimate skill mismatches at an occupational level. This considerable lack of empirical evidence has hampered the design of public policies oriented to reduce skill mismatches, an issue that has been highlighted as one of the leading causes of unemployment and informality in countries like Colombia (Arango and Hamann 2013; Álvarez and Hofstetter 2014; OECD 2015a).

In this respect, the main empirical contribution of this document is a detailed and original analysis of unmet Colombian labour demand, as well as the determination of skill shortages based on novel sources of information, such as job portals and the use of household surveys with an updated occupational classification (ISCO-08) (the occupational information of household surveys was updated thanks to the methodologies developed in this book). Moreover, this study sheds light on the validity of information from job portals for economic analysis, such as the general structure and dynamics of labour demand in Colombia, while providing a method to estimate occupations in shortage.

In particular, the labour demand analysis based on job portals has shown that 1) information collected from job portals is representative of a considerable set of occupations from 2016 to 2018: formal, non-agricultural, non-governmental, non-military, and non-self-employed ("business owners"); and 2) even if vacancy data do not capture a considerable share of some occupations, such as agricultural jobs, the relatively few observations in the database for these occupations might provide insights about new skill requirements and general trends to policymakers, education and training providers, and job seekers (Chapter 8).

Regarding the composition of Colombian labour demand, the analysis shows that: 1) most job positions require a person with at least a high school diploma; 2) in accordance with the previous result, most occupations demanded in Colombia correspond to middle- and low-skilled occupations (such as "Sales demonstrators" and "Kitchen helpers," respectively); 3) job portals are a rich source of information to keep updated Colombian occupational classifications according to changes in the domestic labour market. The most relevant new or

specific job titles found in the vacancy database include "TAT vendors," "CNC operators," and "Baristas" ("new" or "specific" job titles can refer to new job titles or job titles that the ISCO list of occupational titles adapted to Colombia did not previously identify).

Regarding information on skills, the vacancy database shows that 4) the most demanded skills are customer service (knowledge), communication (knowledge), and work in teams (competence); and 5) it is possible to identify new or specific skills in the Colombian labour market (such as "Fintech," "Mailings," and "*Perifoneos*," among others). Thus, it is possible to monitor the changes and specific requirements of the domestic labour market at a low cost by using information from job portals. This single vacancy database enables the analysis of job attributes (occupations demanded) and the skills employers want their workers to have (Chapter 7). Consequently, job portals provide detailed and valuable information about the Colombian labour demand, which was not possible to obtain before via other sources of information (i.e. household surveys).

One of the most distinctive elements of this book is that it has conducted, for the first time in Colombia, a homologated analysis of labour demand and supply information at an occupational level. Specifically, the GEIH has shown that 1) unemployment and informality are widespread phenomena; 2) informal labour (once compared with the formal and unemployed population) tends to be composed of adults over 29 years old, with an education level of high school or less. Consequently, informality rates are higher in low-skilled occupations. 3) In contrast, the unemployed population tends to be characterised by young adults (less than 29 years old), and high- and low-skilled occupations have the highest unemployment rates and prolonged unemployment periods. 4) The labour supply trend analysis has demonstrated that Colombian employment conditions have deteriorated since 2016; and 5) some segments show signs of skill shortages. Indeed, with the use of vacancy and labour supply information, it was found that 30 occupations are currently in short supply, of which 46.7% of the categories belong to high-skilled occupations, while 36.6% and 16.7% correspond to middle- and low-skilled occupations, respectively. This evidence suggests that the formal labour market needs people at all skill levels. In addition, 6) skill mismatch results for Colombia confirm a global trend where occupations related to data, networks, and web professionals show clear signs

of shortage. 7) A detailed analysis of vacancy descriptions can reveal the most important skills in demand for a particular occupation. For instance, SQL, database, and JavaScript are the most demanded skills for "Web and multimedia developers." 8) Moreover, the vacancy analysis has shown that (for instance) for "Web and multimedia developers," object-oriented modelling, creating software design, and Apache Tomcat, among other skills, are becoming more relevant at the time of applying for a job (Chapter 9).

As previously noted, interdisciplinary studies have used online job vacancy data to provide insights about labour demand in different countries. Most of those studies do not properly discuss representativeness issues (which might affect data results) or do not combine and analyse labour supply data and information from job portals to estimate possible skill mismatches and skill requirements. The most important ongoing project similar to this document (in term of objectives) is the "Big Data analysis from online vacancies," conducted by Cedefop (see Chapter 4). However, even compared to the Cedefop project, this book has produced different contributions, which can be summarised as follows.

a. This study has focused on investigating the advantages, limitations, and uses of information from job portals for Colombia, which is a non-European developing country with severe skill mismatch issues;
b. introduced a theoretical framework regarding labour market mismatches and the potential usefulness of job portals to tackle those phenomena in the Colombian context (Chapters 2 to 4);
c. discussed and proposed methods to collect and process a wider number of variables (e.g. education, wages, etc.) (Chapters 5 and 6);
d. suggested a new mixed-methods approach to classify job titles into occupations and to identify skills for a country that does not have official skills dictionaries (Chapter 6);
e. analysed a wider scope of variables, such as educational requirements, wages, sector, among others (Chapter 7);
f. used a more extended period of data study for Colombia (January 2016-ongoing) compared to Cedefop (April 2018-ongoing). This extended period enables the analysis of labour market trends and seasons in Colombia (Chapter 7);

g. provided a framework and tested the validity and consistency of information from job portals (Chapter 8);
h. combined job portal and household survey data to determine skill shortages in Colombia (Chapter 9).

In conclusion, this book has contributed to the development of a conceptual and methodological framework that enables the generation and robust analysis of much needed empirical data, such as those regarding skill requirements, as well as the estimation of skill shortages at an occupational level. Moreover, the empirical contributions of this study include evidencing that (Big Data) information from job portals can complement traditional data (e.g. household or employer surveys) for a consistent, comprehensive, and fruitful labour market analysis to support public policy advice. Therefore, this book has various important implications for national statistics offices, policymakers (e.g. ministries), education and training providers, and career advisers.

10.5. Implications for practice and policy

As mentioned above, one of the most important contributions of this study is its significance for national statistics offices, policymakers (e.g. Ministry of Labour and Education), education and training providers, and career advisers, which are analysed in more detail in the following subsections.

10.5.1. For national statistics offices

The importance of this research for national statistics offices consists of demonstrating that, with the adequate techniques, online information (in this case from job portals) can be an important source of data that can complement the statistical analysis of data collected using “traditional” methods. However, it is necessary to implement novel techniques to test and use these sources of information. Thus, the first specific input for the offices of national statistics is that the faster they adopt new techniques, the better they can benefit from the abundant information produced online to fill information gaps.

As mentioned in Chapter 4, despite the fact that vacancy information is not being created for economic analysis, this source has proved to offer consistent data regarding the characteristics of the Colombian labour market. However, the scope of information that can be extracted from job portals depends on the research focus. For instance, this book has demonstrated that it is possible to accurately determine skill shortages via vacancy information. Nonetheless, national statistics offices need to create their analytical frameworks to determine whether information from job portals (among other sources) can be used to answer other economic questions. Thus, the second input of this document is that it urges national statistics offices worldwide to debate and determine the scope of information from job portals based on national contexts.

As highlighted in this research, vacancy data have an advantage when compared to other "traditional" sources of labour market information given that job portals provide real-time, detailed, and accurate information about economic seasons and trends at an occupational level (as they did for Colombia and would potentially do for other countries too). Thus, other debate should focus on using vacancy time series for measuring labour demand seasons and trends, and testing whether the vacancy database is an accurate source for the early identification of economic cycles. To do so, it is paramount to continue collecting vacancy information in a consistent manner. For instance, as mentioned in Chapter 4, other countries, such as the US and Australia, have developed vacancy indexes based on online information to provide short-term measures of labour demand at different disaggregation levels. Furthermore, this debate should also focus on how detailed information from job portals, such as skills and experience, among other employer requirements, can help economic agents to make informed decisions. Vacancy information can serve different purposes (e.g. skills, educational, regional, structural, trend analysis, etc.) and offices for national statistics can assist with determining the validity of each potential use for vacancy information.

Third, as discussed in Chapter 7, some countries might identify emerging occupations and skills (e.g. O*NET or ESCO) faster than other countries because they have relatively higher budgets to conduct employer surveys among other continuous efforts. However, offices for national statistics can start looking at information from job portals immediately to identify occupational and skill changes rapidly (see Subsection 10.7.2). As demonstrated in Chapters 6 to 9,

based on online vacancy information, it is possible to build robust text mining and classification methods (e.g. machine learning) to regularly identify and include new job titles, skills, and occupations into occupational classifications at a low cost.

Moreover, this book has highlighted the importance of updating and continuously adapting occupational classifications drawn from household surveys. As discussed earlier, the usage of outdated classifications (such as the 1970 SOC) might lead to misclassification and/or underestimation/overestimation of certain occupational categories. Given that obsolete classifications make internal and international comparisons difficult, the DANE should endeavour to update its occupational classifications.

10.5.2. For policymakers

The main implication for policymakers is that they can use information from job portals along with traditional data to create a set of coherent public policies that tackle skill mismatches. In general, with the implementation of the methods and results of this study, the government is able to inform education and training providers and job seekers about the most demanded skills and occupations in skill mismatch. With this action, the government has the opportunity to reduce imperfect labour market information, and, hence, interested parties can make better informed employment decisions (Chapter 2).

More specifically, the Colombian Public Employment Service (UAESPE, for its acronym in Spanish) could take advantage of these methods to develop a profiling framework based on the statistical model presented in the previous chapters. As demonstrated in these chapters, the profile of people in the informal economy is different from those who are unemployed. Furthermore, some occupations show clear signs of skill shortages. Consequently, given an informal or unemployed person's occupation (among other characteristics), it is possible to know which are the most similar job(s) to this person's occupation that have skill shortages. Based on people's profiles and the identification of occupations in shortage, the UAESPE could effectively assist informal and unemployed individuals to find the best and shortest route to obtain a formal job.

As mentioned in Chapter 3, one of the reasons that might explain skill mismatches is the relatively high presence of educational and job training

programmes that are not aligned with employer requirements and have a low standard of quality. The Ministry of Labour and the Ministry of Education could encourage education and training providers to increase their courses or degrees related to those occupations in skill mismatch. For instance, Chapter 9 has shown that "Electrical line installers and repairers" and "Structural-metal preparers and erectors" display strong signs of skill mismatch (see Table 9.9). Thus, with the training programme-occupation matrix created by the Ministry of Education and the SENA, it is possible to determine which courses should be increased or improved. In the case of Colombia, the matrix indicates that a required programme for "Electrical line installers and repairers" is "installation of telecommunications services, installation and maintenance of HFC networks," while "Structural-metal preparers and erectors" require knowledge of "construction of concrete structures" and "light constructions." Consequently, such training programmes should be encouraged.

Related to the above point, ministries could encourage education and training providers to adapt their curricula to the skills identified in vacancy announcements. As discussed earlier, skills are a crucial factor to find a job. Furthermore, for a considerable number of occupations, the vacancy information allows identifying, over a short period, the most relevant skills in demand. Consequently, keeping up-to-date the educational and training supply according to skill requirements is an important step to avoid future increases in skill shortages, and also to decrease the current incidence of unemployment and informality in the labour market due to imperfect information.

Currently, Colombia is building its national qualifications framework (NQF).[149] One of the most important inputs for the NQF is a detailed labour demand analysis, given that the study of labour demand allows identifying the qualifications and skills (competences) related to each occupation. Therefore, the vacancy database and methodologies presented in this book could serve as a guide to the Ministry of Education (among other institutions) regarding

---

149 An NQF "describes the qualifications of an education and training system and how they interlink. National qualifications frameworks describe what learners should know, understand, and be able to do on the basis of a given qualification. These frameworks also show how learners can move from one qualification, or qualification level, to another within a system" (Quality and Qualifications Ireland, n.d.).

how vacancy information might be used to profile occupations and identify qualifications and skills in demand.

Some European countries, such as Austria, Germany, and the UK, use National Occupational Standards (NOS) or occupational profiles as primary units to identify skills and guide the construction of vocational qualifications (EQF Predict 2008). As this document has evidenced, detailed information and analysis based on data collected from job portals can provide insights for NOS and for occupational profiles, which in turn aids the construction of qualifications needed in the NQF of each country.

In countries, such as the UK, US, and New Zealand, there exist initiatives to make labour market data available to the public (Hughes, Bimrose, and Barnes, n.d.). For instance, the "Labour Market Information (LMI) for All" initiative in the UK has important added values for the design of well-oriented public policies given that this project combines and standardises labour market data from various sources (e.g. the Labour Force Survey and the Annual Survey of Hours and Earnings carried out by the Office for National Statistics, and the Employer Skills Survey (ESS) performed by the UK Commission for Employment and Skills), the purpose of which is to provide high quality and reliable information at an occupational level to orient career paths.

One attempt of this kind of initiative in Colombia is FILCO (Fuente de Información Laboral de Colombia) undertaken by the Ministry of Labour. However, the platform only gathers information from sources such as the DANE, it utilises aggregated labour analysis and is not user-friendly. Consequently, the information provided through this platform is insufficient to assist people with making career and curriculum decisions. The Ministry of Labour can improve the services offered through FILCO by integrating labour market analysis and methods like the ones described in this book.

Moreover, the government needs to promote the use of initiatives such as the FILCO by providing a user-friendly tool with updated, disaggregated, robust, and relevant labour market information, open to the public. An example of such open vacancy database is "skills-OVATE: Skills Online Vacancy Analysis Tool for Europe" launched by the Cedefop[150] in 2019 (Chapter 4). This

150 See https://www.cedefop.europa.eu/en/data-visualisations/skills-online-vacancies/skill-sets-occupations.

kind of tools can be very useful for public and private institutions in order to provide better training and career advice, among other services.

Finally, given the dynamic nature of the labour market, occupations that are in skill shortage today might not have this problem in the future. Consequently, the quantitative approach for occupational matching and skills profiles (described in this study) needs to be updated at least monthly to monitor trends, seasons, and potential cycles of the labour dynamics over the short term. Moreover, the use of vacancy information for other purposes than the identification of skill shortages will involve different institutions, such as Offices for National Statistics, ministries, etc. Government institutions, such as the DANE, the Public Employment Service, and the Ministry of Labour, need to work together in order to update this methodology, which will allow improving public policies related to a better management of human resources.

10.5.3. For education and training providers

As mentioned in Chapter 2, the match between employers and job seekers depends substantially on how education and training systems answer and adapt to company requirements. However, in Colombia, educational and job training programmes are not aligned with employer demands, and programmes with low standards of quality have proliferated. This phenomenon is partly due to the lack of an articulated human capital formation system with accurate tools to address education and job training programs (Chapter 3). This document has demonstrated that, with the proper methods, job portals are a novel and valuable tool that allow identifying current occupational and skill requirements in the Colombian labour market over time (Chapters 7 to 9). Based on these insights, educational and job training can provide appropriate skills to prepare people for formal jobs. Consequently, this book has implications regarding how education and training providers can make use of vacancy information. It is expected that one of the main concerns of education and training providers is to provide relevant curricula. In this way, people studying in relevant programmes will find it easier to get a formal job, institutions offering relevant courses will gain popularity, and thus the number of enrolled individuals or the willingness to pay to be enrolled in these institutions will increase (which, for the institutions, might represent higher profits).

Thus, education and training providers need to consider employer requirements when planning their curricula. In cases where the government is not able to provide updated labour demand information, private institutions can create a system to monitor the segment of the labour market in which they are interested. This book has demonstrated that it is possible to develop, at a relatively low cost, a system that monitors company requirements for a significant number of occupations or a particular segment of the labour market. The main concepts behind this system have been analysed and defined in this document. Consequently, any institution or person with an interest in building a particular labour demand monitoring system will find helpful the discussions and findings of this book.

10.5.4. For career advisers

Career advisers can improve their efficiency by considering the analysis of data collected from job portals to inform people's decisions regarding their educational, training, and work options. This document has shown that, despite overall socio-economic improvements during the last decades, employment conditions have deteriorated recently. Moreover, labour demand has a dynamic nature, and some occupations and skills emerge while others decline. As discussed in Chapter 2, the more efficiently information and advice is provided to job seekers, the better the labour market outcomes. Consequently, career advisers should use and, if possible, carry out analyses of information from job portals to provide better insights to job seekers.

Given the results from the vacancy data, career advisers can offer accurate information on occupations in demand, as well as on educational and training programmes available in a certain region. Importantly, these institutions can help people to make a connection between education, training, and occupations demanded, and inform people regarding the costs and benefits of a specific career path. For instance, using vacancy information, it is possible to know the average salary for an occupation, while information regarding the costs and duration of a particular educational and training program is usually available in each educational institution or the Ministry of Education (Sistema Nacional de Información de Educación Superior, SNIES). Thus, career advisers are able to consistently estimate and inform people about the returns of a particular career path.

Furthermore, personal guidance services provided by career advisers can be improved because the vacancy database gives insights about the most critical sector-specific, cross-specific, and transversal skills for an occupation (Chapters 7 to 9). Therefore, career advisers have a proper tool to determine what are the most pertinent skill training programmes for a person with specific characteristics and vocational aspirations. Finally, using vacancy information and the results of this book, career advisers can potentially direct job seekers at a regional level to companies that currently have job vacancies, and assist people with preparing their CVs according to employer requirements. Consequently, integrating the vacancy analysis presented in this book with initiatives such as the LMI for the UK and the FILCO for Colombia can improve the efficiency of career advisers who could provide advice based on proper, continuously updated, and publicly available labour supply and demand information.

10.6. Limitations

Despite its many contributions to advancing our understanding of the topic, this research is not exempt from limitations. The sources of these limitations originate from the type of information available in online job vacancies, the methods for collecting labour demand data, the relatively short period used for the analysis (which will be resolved as web scraping continues), the lack of official external information for the comparison of results, and other issues such as multicollinearity, omitted variables, etc., in the estimated regressions.

Although it has been argued that it is not necessary to know the precise amount of vacancies in the economy to identify possible skill shortages (Chapter 8), having a rough estimation of the number of vacancies in the country might be helpful to tackle skill mismatches and their consequences. For instance, it could be helpful if education and training providers have an idea of the number of courses and people they will train in a specific occupation; however, at this moment, it is not possible to determine the exact number of vacancies available in the Colombian economy, mainly due to the absence of a vacancy census or similar tool. Thus, one of the specific limitations of this book is that it identifies skill mismatches but does not provide the number of occupational shortages.

Related to the above point, in terms of data collection, it is essential to recognise that (with the techniques available today) there is no way to demonstrate that all duplicated observations have been dropped. However, Chapter 8 has shown that latent duplication issues do not considerably affect the validity of the vacancy database.

Moreover, there are gaps or weaknesses in the vacancy information content. This document has identified the most demanded skills by Colombian employers. Nevertheless, nothing is said about the level and extent of the required skills. For instance, an employer might need “English” for a certain vacancy. With this limited information, it is not possible to know whether employers are asking for an advanced or intermediate level of English, or whether they are referring to speaking, writing, or listening skills in English. In most cases, employers do not provide enough information to determine the level and extent of the required skills. Thus, in this case, for an extensive analysis of the skills being demanded, it is necessary to complement the vacancy information (whenever it is possible and appropriate) with employer in-depth surveys.

Although Chapter 6 has shown that it is plausible to impute missing variables such as wages, for some variables, at this moment, it is not possible to properly apply imputation methods. For instance, as discussed earlier, due to the high participation of “Temporary employment agency activities” and missing values in the sector variable, it is not feasible to construct an accurate analysis of the labour market by sector. The issue of missing values in some variables is an important limitation for the analysis of skill mismatches.

Finally, the regression conducted in Chapters 8 and 9 can have potential multicollinearity, omitted variables, and other issues. On the one hand, it is important to note that, although the vacancy data are rich in information, this source might not have all the variables (omitted variables) to make a proper wage regression. Variables such as the type of contract, number of hours required to work, and other relevant variables were not considered because the vacancy database does not have these kinds of information or the proportion of missing values is considerably high. On the other hand, the estimations with the GEIH, for instance, might not control for the unobserved heterogeneity of individuals. Unobservable skill differentials (e.g. self-motivation) and preferences might affect labour market outcomes, such as unemployment, informality or wage

levels (Heckman, Stixrud, and Urzua 2006). However, further examination is required to address these problems properly.

10.7. Further research

This book has demonstrated that it is possible to build a robust theoretical and methodological framework to collect and analyse vacancy data collected from job portals to tackle skill mismatch issues. This robust framework brings statistical confidence to using the results of information from job portals for different purposes. Based on the main findings and limitations discussed earlier, online vacancy data in Colombia can, potentially, be improved by refining text mining algorithms, identifying new occupations, and making international comparisons; as a result, they can be used for academic or public policy purposes. Next, this section examines these main future research directions in more detail.

10.7.1. Improving machine learning and text mining algorithms

As discussed in Chapter 6, a considerable percentage of non-coded job titles were due to the absence of key information in the job title variable. In many cases, the words in job titles without an occupational code did not provide adequate information regarding the job position; for instance, a regular word with these characteristics is "*Bachilleres*" (which in English means "Undergraduate"). Based on this limited information, neither automatic nor manual classifiers can assign an occupational code to an observation with these characteristics. Similarly, there is a portion of the information that has no ISIC code because the company name variable was not enough to identify the corresponding industrial code.

One reasonable alternative to overcome these issues is considering the job description. Sometimes, information about the job position or a company's activities is in the job description rather than in the job title or in the company's name. Thus, processing and identifying specific patterns in the job description might increase the number of observations with an occupational and industrial code. However, to carry out this task, it is necessary to develop an advanced text mining method that recognises different linguistic patterns used

by employers to describe an occupation and the activities of a company in the vacancy description. It is important to note that despite algorithm improvements, observations might remain without sufficient or clear information to be able to assign an occupational or industrial code.

10.7.2. New job titles and potential new occupations

Chapter 7 has shown that job portals provide updated information regarding new job titles required by employers, such as "TAT vendors" and "Picking and packing assistants." In some cases, the new job titles are already listed in different versions of ISCO-08 in other countries, as is the case for "CNC operators" or "Bobcat operators." In these cases, vacancy information might help to identify and update similar job titles demanded by Colombian companies that are not included in the Colombian ISCO-08 classifications but are listed in other countries' versions.

However, in some cases, new Colombian job titles are not listed in other international versions of ISCO-08. In these cases, it is necessary to evaluate whether a certain new job title corresponds to a new occupation or, on the contrary, the new job title can be assigned to an existing occupational ISCO-08 category.

One of the most complete systems for the identification of "new and emerging (N&E) occupations" is the O*NET in the US. Developing a system like the O*NET for Colombia would be costly in terms of time and money, since it is necessary to make agreements with different institutions and obtain enough budget to conduct in-depth interviews in each sector and different occupations, among other elements. Despite the high costs of the O*NET, this system provides a sufficiently solid theoretical and methodological framework to design a methodology (based on vacancy information and other available data) that identifies new and emerging occupations in Colombia.

Following the O*NET definitions, an N&E occupation is defined as follows:

- "The occupation involves significantly different work than that performed by job incumbents of other occupations, as determined by NC State and O*NET research consultants; and
- The occupation is not adequately reflected by the existing O*NET-SOC structure" (O*NET 2006).

Based on these definitions, it is possible to identify new occupations. For instance, one of the most frequent new job titles identified in Colombian job portals was “TAT vendors.” The question is whether that job title should be considered as a new occupation or a new title of an existing occupational category. One way to address this issue at a low cost and over a short time period is by using the vacancy database. As shown in Chapters 7 and 9, online vacancy data properly capture demanded skills (among other requirements) by occupation and potentially by job titles.

Consequently, vacancy information can, potentially, determine whether the set skills and tasks required for an occupation such as “TAT vendors” are different from other types of sellers. With a list of potential new occupations, activities, and skills, institutions in charge of adopting and updating the national occupational classification system can be more efficient in these tasks by carrying out in-depth interviews with experts and focusing on potential new occupations. This area of future research would aim to determine to what extent and how information from job portals can provide a quick and inexpensive way to keep updated and adapt occupational classifications according to national contexts.

10.7.3. International comparison

Over the last decades, there has been a skill-biased technological change, which has increased labour demand and wages for skilled labour compared with unskilled labour (Autor, Katz, and Krueger 1998). However, countries have not implemented the same technological changes due to differences in the supply of skills, economic cycles, and existing national regulations (Acemoglu 1998; Pertold-Gebicka 2014). As Acemoglu and Zilibotti (2001) point out, wealthier economies tend to employ more skilled workers, creating skill complementarities and increasing the productivity of those regions. Consequently, changes in preferences for skilled workers (labour demand for skills) have enlarged the gap in productivity and wages between poor and wealthy countries (Acemoglu 1998; Broecke 2016).

Since the information for labour demand is scarce or imprecise, it is difficult to analyse and compare company requirements in different countries (Handel 2012; Kureková, Beblavy, and Thum, 2014; Reimsbach-Kounatze 2015; Tijdens,

Beblavy, and Thum-Thysen 2015). Hence, the purpose of future international research could be to elaborate a standardised approach to collecting vacancy information, which would allow international comparisons of unmet labour demand and an analysis of differences between various regions in the world, mainly in terms of occupations and skills demanded by job portals. These sources of information can provide insights to identify different technological paths (such as job polarisation or skill traps) (Carnevale, Jayasundera, and Repnikov 2014). Consequently, the analysis of unsatisfied labour demand could provide answers regarding how far or distinct developing economies are from more developed economies in terms of their labour demand for skills.

10.8. Conclusions

This book has investigated to what extent a web-based model of skill mismatches can be developed for Colombia, where information regarding labour demand is scarce. It was found that online job portals can provide high quality, real-time, and detailed data to decrease imperfect information in the labour market and to tackle skill mismatch issues. However, before using job portal data for economic analysis, it is necessary to undertake an exhaustive and continuous evaluation of these sources of information.

The evidence for the Colombian case suggests that for a considerable set of occupations, information from job portals is representative of the urban unmet labour demand. This information provides abundant, relevant, and consistent insights regarding the skills demanded by employers over time. Consequently, the vacancy database—along with the Colombian household survey—can be used to create a quantitative system to identify skill mismatches and skill requirements. The evidence from this system has shown that there is a set of occupations at different skill levels that experience skill shortages, and the profile of people in the informal economy is different from that of unemployed people. Thus, it is possible to design better public policies according to employer requirements and different profiles of people outside of the formal labour market.

The findings of this book make important conceptual, methodological, and empirical contributions as they demonstrate that (with the proper techniques) information collected from job portals can fulfil the conceptual requirements to

be considered high-quality data for labour market analysis; develop a detailed framework and methods to collect, clean, and organise vacancy data; test the internal and external validity of data from this source of information; provide a detailed and consistent labour market analysis that reveals relevant and previously unknown characteristics of the Colombian labour demand; and show the advantages and limitations of a web-based model of skill mismatches adapted for Colombia. Furthermore, other countries, especially those with similar characteristics to the Colombian economy (high unemployment and informality rates and scarce information for labour demand) can benefit from adopting a web-based model of skill mismatches (skill shortages) based on the contributions of this book. In this regard, whilst this document has advanced current understandings of the topic, it also opens new avenues of enquiry for future research.





References

Acemoglu, Daron. 1998. “Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality.” *The Quarterly Journal of Economics* 113, no. 4: 1055–89.

Acemoglu, Daron, and David Autor. 2011. “Skills, Tasks and Technologies: Implications for Employment and Earnings.” In *Handbook of Labor Economics*, Volume 4B, edited by David Card and Orley Ashenfelter, 1043–171. San Diego, CA: Elsevier.

Acemoglu, Daron, and Fabrizio Zilibotti. 2001. “Productivity Differences.” *The Quarterly Journal of Economics* 116, no. 2: 563–606.

Adalet McGowan, Müge, and Dan Andrews. 2015. *Skill Mismatch and Public Policy in OECD Countries*. OECD Economics Department Working Papers, No. 1210. Paris: OECD Publishing.

Aguilar, Luis Joyanes. 2016. *Big Data. Análisis de grandes volúmenes de datos en organizaciones*. 1st ed. Mexico, D.F.: Alfaomega Grupo Editor.

Albrecht, James, Lucas Navarro, and Susan Vroman. 2007. “The Effects of Labour Market Policies in an Economy with an Informal Sector.” *The Economic Journal* 119, no. 539: 1105–29.

Alexa. 2017. “Website Traffic.” Accessed October 15, 2017. https://www.alexa.com/siteinfo.

Allen, Jim, Mark Levels, and Rolf van der Velden. 2013. *Skill Mismatch and Skill Use in Developed Countries: Evidence from the PIAAC Study*. ROA Research Memorandum 017, Maastricht University: Research Centre for Education and the Labour Market (ROA).

Almeida, Rita, Jere Behrman, and David Robalino. 2012. *The Right Skills for the Job? Rethinking Training Policies for Workers*. Washington, DC: The World Bank.

Álvarez, Andrés, and Marc Hofstetter. 2014. “Job Vacancies in Colombia: 1976--2012.” *IZA Journal of Labor & Development* 3, no. 1: 1–11.

Andrews, Martyn J., Steve Bradley, Dave Stott, and Richard Upward. 2008. "Successful Employer Search? An Empirical Analysis of Vacancy Duration Using Micro Data." *Economica* 75, no. 299: 455–80.

Antenucci, Dolan, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D. Shapiro. 2014. *Using Social Media to Measure Labor Market Flows*. NBER Working Paper 20010. Cambridge, MA: National Bureau of Economic Research.

Arango, Luis Eduardo, and Franz Alonso Hamann. 2013. *El mercado de trabajo en Colombia: hechos, tendencias e instituciones*. 1st ed. Bogotá: Banco de la República.

Arrow, Kenneth Joseph. 1962. "The Economic Implications of Learning by Doing." *The Review of Economic Studies* 29, no. 3: 155–73.

Askitas, Nikolaos, and Klaus F. Zimmermann. 2009. *Google Econometrics and Unemployment Forecasting*. Discussion Paper No. 4201. Bonn, Germany: IZA.

____. 2015. "The Internet as a Data Source for Advancement in Social Sciences." *International Journal of Manpower* 36, no. 1: 2–12.

Asplund, Rita. 2005. *The Provision and Effects of Company Training: A Brief Review of the Literature. Nordic Journal of Political Economy* 31: 47–73.

Attewell, Paul. 1990. "What Is Skill?" *Work and Occupations* 17, no. 4: 422–48.

Australian Government. 2018. "Internet Vacancy Index." Accessed August 20, 2018. https://data.gov.au/dataset/internet-vacancy-index.

Autor, David H. 2001. "Wiring the Labor Market." *The Journal of Economic Perspectives* 15, no. 1: 25–40.

Autor, David H., and David Dorn. 2012. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103, no. 5: 1553–97.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. "Computing Inequality: Have Computers Changed the Labor Market?" *The Quarterly Journal of Economics* 113, no. 4: 1169–213.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2006. *The Polarization of the U.S. Labor Market*. NBER Working Paper No. 11986. Cambridge, MA: National Bureau of Economic Research.

Azzone, Giovanni. 2018. "Big Data and Public Policies: Opportunities and Challenges." *Statistics & Probability Letters* 136: 116–20.

Backhaus, Kristin B. 2004. "An Exploration of Corporate Recruitment Descriptions on Monster. Com." *The Journal of Business Communication (1973)* 41, no. 2: 115–36.

Bahk, Byong, and Michael Gort. 1993. “Decomposing Learning by Doing in New Plants.” *Journal of Political Economy* 101, no. 4: 561–83.

Banfi, Stefano, and Benjamín Villena-Roldán. 2019. “Do High-Wage Jobs Attract More Applicants? Directed Search Evidence from the Online Labor Market.” *Journal of Labor Economics* 37, no. 3: 715–46.

Barnichon, Regis. 2010. “Building a Composite Help-Wanted Index.” *Economics Letters* 109, no. 3: 175–78.

Barrett, Alan, and Philip J. O’Connell. 1999. “Does Training Generally Work? The Returns to in-Company Training.” *ILR Review* 54, no. 3: 647–62.

Bassanini, Andrea, Alison L. Booth, Giorgio Brunello, Maria De Paola, and Edwin Leuven. 2007. *Workplace Training in Europe*. Discussion Paper No. 1640. Bonn, Germany: IZA.

BBVA. 2018. “The Five V’s of Big Data.” Accessed May 5, 2018. https://www.bbva.com/en/five-vs-big-data/.

Becker, Gary S. 1962. “Investment in Human Capital: A Theoretical Analysis.” *Journal of Political Economy* 70, no. 5, Part 2: 9–49.

____. 1994. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. 3rd ed. Chicago: The University of Chicago Press.

Bell, Linda A. 1997. *The Impact of Minimum Wages in Mexico and Colombia*. Washington, DC: The World Bank.

Belloni, Michele, Agar Brugiavini, Elena Meschi, and K. G. Tijdens. 2014. *Measurement Error in Occupational Coding: An Analysis on SHARE Data*. Deptartment of Economics Working Paper Series No. 24/WP/2014. Venice, Italy: University Ca’ Foscari of Venice.

Bernal S., Raquel. 2009. “The Informal Labor Market in Colombia: Identification and Characterization.” *Revista Desarrollo y Sociedad* 63: 145–208.

Bethmann, Arne, Malte Schierholz, Knut Wenzig, and Markus Zielonka. 2014. “Automatic Coding of Occupations.” *Proceedings of Statistics Canada Symposium 2014: Beyond Traditional Survey Taking: Adapting to a Changing World*. Hull, Quebec, Canada.

Black, Sandra E., and Lisa M. Lynch. 1995. *Beyond the Incidence of Training: Evidence from a National Employers Survey*. NBER Working Paper No. 5231. Cambridge, MA: National Bureau of Economic Research.

Blanchard, Olivier Jean, and Peter Diamond. 1989. “The Beveridge Curve.” *Brookings Papers on Economic Activity* 1: 1–76.

Bleakley, Hoyt, and Jeffrey C. Fuhrer. 1997. “Shifts in the Beveridge Curve, Job Matching, and Labor Market Dynamics.” *New England Economic Review* 28: 3–19.

Blundell, Richard, Lorraine Dearden, Costas Meghir, and Barbara Sianesi. 1999. “Human Capital Investment: The Returns from Education and Training to the Individual, the Firm and the Economy.” *Fiscal Studies* 20, no. 1: 1–23.

Booz, Michael. 2018. “These Are the 5 Types of Jobs with the Most Turnover.” *LinkedIn Talent Blog*. Last modified April 12, 2018. https://business.linkedin.com/talent-solutions/blog/talent-analytics/2018/these-are-the-5-types-of-jobs-with-the-most-turnover.

Bosworth, Derek. 1993. “Skill Shortages in Britain.” *Scottish Journal of Political Economy* 40, no. 3: 241–71.

Bosworth, Derek L., Peter Dawkins, and Thorsten Stromback. 1996. *The Economics of the Labour Market*. Harlow: Longman.

Broecke, Stijn. 2016. “Do Skills Matter for Wage Inequality?” *IZA World of Labor*. doi: 10.15185/izawol.232

Brunello, Giorgio, and Martin Schlotter. 2011. “Non-Cognitive Skills and Personality Traits: Labour Market Relevance and Their Development in Education & Training Systems.” IZA Discussion Papers No. 5743. Bonn, Germany: IZA.

Burdett, Ken, and Eric Smith. 2002. “The Low Skill Trap.” *European Economic Review* 46, no. 8: 1439–51.

Burning Glass Technologies. 2017. *The Digital Edge: Middle-Skill Workers and Careers*. https://www.burning-glass.com/wp-content/uploads/Digital_Edge_report_2017_final.pdf

Cabrera, Armando, Dora Rodríguez, Fernando Vargas, Álvaro Barragán, Esperanza Rubiano, and Camilo Cifuentes. 1997. “Clasificación nacional de ocupaciones.” Bogotá: SENA.

Cahuc, Pierre, Stéphane Carcillo, and André Zylberberg. 2014. *Labor Economics*. Cambridge, MA: MIT Press.

Cambridge Econometrics. 2013. *Assumptions for the Baseline and ‘Smart Efficiency and Growth’ Scenarios for Worcestershire Districts*. Cambridge, UK: Cambridge Econometrics.

Cappelli, Peter H. 2015. “Skill Gaps, Skill Shortages, and Skill Mismatches: Evidence and Arguments for the United States.” *ILR Review* 68, no. 2: 251–90.

Carnevale, Anthony P., Tamara Jayasundera, and Dmitri Repnikov. 2014. *Understanding Online Job Ads Data. A Technical Report.* Washington, DC: Georgetown University.

Cisco. 2017. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2016-2021.* San Francisco: Cisco.

Cohen, William, Pradeep Ravikumar, and Stephen Fienberg. 2003. "A Comparison of String Metrics for Matching Names and Records." *Proceedings of the KDD2003.* https://www.cs.cmu.edu/afs/cs/Web/People/wcohen/postscript/kdd-2003-match-ws.pdf

Comisión Económica para América Latina y el Caribe (CEPAL). 2016. *Estado de la banda ancha en América Latina y el Caribe 2016*. Last modified October 2016. https://repositorio.cepal.org/bitstream/handle/11362/40528/6/S1601049_es.pdf.

Conference Board. n.d. "The Conference Board Help Wanted OnLine® (HWOL)." Accessed August 20, 2018. https://www.conference-board.org/data/helpwantedonline.cfm.

Consejo Nacional de Política Económica y Social (CONPES). 2010. *Lineamientos de Política Para El Fortalecimiento Del Sistema de Formación de Capital Humano SFCH.* Bogotá: Departamento Nacional de Planeación.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97, 2: 31–47.

Cunningham, Wendy, and Paula Villaseñor. 2016. *Employer Voices, Employer Demands, and Implications for Public Skills Development Policy.* Washington, DC: The World Bank.

Departamento Administrativo Nacional de Estadística (DANE). 2009. *Metodología Gran Encuesta Integrada de Hogares.* Bogotá: DANE.

____. 2014. *Encuesta de formación de capital humano*. Last modified April 30, 2014. https://www.dane.gov.co/index.php/estadisticas-por-tema/industria/encuesta-de-formacion-de-capital-humano.

____. 2015. *Clasificación Internacional Uniforme de Ocupaciones. Adaptada para Colombia*. Last modified July 21, 2015. https://www.dane.gov.co/files/sen/nomenclatura/ciuo/CIUO_08_AC_2015_07_21.pdf.

____. 2017a. "Empleo informal y seguridad social." Accessed January 27, 2017. http://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-informal-y-seguridad-social.

____. 2018a. *Producto Interno Bruto (PIB) Departamental. 2017 Preliminar*. Last modified June 29, 2018. https://www.dane.gov.co/files/investigaciones/pib/departamentales/B_2015/Bol_dptal_2017preliminar.pdf

____. 2018b. "Human Capital Formation Survey." Accessed July 9, 2018. https://www.dane.gov.co/index.php/en/statistics-by-topic-1/education/human-capital-formation-survey.

____. 2018c. *Informe de gestión DANE-FONDANE*. Accessed July 9, 2018. https://www.dane.gov.co/files/control_participacion/rendicion_cuentas/Informe_gestion_2017_DANE.pdf

Dehnbostel, Peter. 2002. "Bringing Work-Related Learning Back to Authentic Work Contexts." In *Transformation of Learning in Education and Training. Key qualifications revisited*, edited by Pekka Kämäräinen, Graham Attwell, and Alan Brown, 190–202. Luxembourg: Office for Official Publications of the European Communities.

Deming, David, and Lisa B. Kahn. 2018. "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals." *Journal of Labor Economics* 36, no. (S1): S337–S369.

Desjardins, Richard, and Kjell Rubenson. 2011. *An Analysis of Skill Mismatch Using Direct Measures of Skills*. OECD Education Working Papers, No. 63. Paris: OECD Publishing.

Dierdorff, Erich C., Jennifer J. Norton, Donald W. Drewes, Christina M. Kroustalis, David Rivkin, and Phil Lewis. 2009. *Greening of the World of Work: Implications for O*NET®-SOC and New and Emerging Occupations*. Raleigh, NC: National Center for O*NET Development.

Dobbin, Kevin K., and Richard M. Simon. 2011. "Optimally Splitting Cases for Training and Testing High Dimensional Classifiers." *BMC Medical Genomics* 4, no. 31.

Doeringer, Peter B., and Michael J. Piore. 1971. *Internal Labor Markets and Manpower Analysis*. 1st ed. London: ME Sharpe.

Edelman, Benjamin. 2012. "Using Internet Data for Economic Research." *Journal of Economic Perspectives* 26: 189–206.

Elsby, Michael W. L., Ryan Michaels, and David Ratner. 2015. "The Beveridge Curve: A Survey." *Journal of Economic Literature* 53, no. 3: 571–630.

EQF Predict. 2008. *Typology of National Occupational Standards (Draft Version)*. Accessed July 9, 2018. https://www.project-predict.eu/fileadmin/Dateien/Workpackages/WP3/Typology_of_occupational_standards_10_08.pdf

European Centre for the Development of Vocational Training (Cedefop). 2010. *The Skill Matching Challenge: Analysing Skill Mismatch and Policy Implications*. Luxembourg: Office for Official Publications of the European Communities.

European Commission. 2015. *Analytical Web Note 7/2015. Measuring Skills Mismatch*. Accessed July 20, 2018. https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=7860.

____. 2012a. *Quantifying Skill Needs in Europe*. 30. Luxembourg: Office for Official Publications of the European Communities.

____. 2012b. *Skill Mismatches*. 21. Luxembourg: Office for Official Publications of the European Communities.

____. 2015. *Skill Shortages and Gaps in European Enterprises: Striking a Balance between Vocational Education and Training and the Labour Market*. 102. Luxembourg: Office for Official Publications of the European Communities.

____. 2018. *Big Data Analysis: Online Vacancies*. Luxembourg: Office for Official Publications of the European Communities.

____. 2019. *Online Job Vacancies and Skills Analysis: A Cedefop Pan-European Approach*. Luxembourg: Office for Official Publications of the European Communities.

____. 2017. *ESCO Handbook. European Skills, Competences, Qualifications and Occupations*. Last modified December 8, 2017. https://op.europa.eu/en/publication-detail/-/publication/ce3a7e56-de27-11e7-a506-01aa75ed71a1/language-en.

Eurostat. 2017. "Job Vacancies." Accessed July 15, 2018. https://ec.europa.eu/eurostat/web/labour-market/job-vacancies

Farm, Ante. 2003. *Defining and Measuring Unmet Labour Demand*. Working Paper Series 1/2003. Stockholm: Swedish Institute for Social Research.

Flórez, Luz Adriana, Leonardo Fabio Morales-Zurita, Daniel Medina, and José Lobo C. 2017. *Labour Flows Across Firm's Size, Economic Sectors and Wages in Colombia: Evidence from Employer-Employee Linked Panel*. Bogotá: Banco de la República de Colombia.

Freije, Samuel. 2002. *Informal Employment in Latin America and the Caribbean: Causes, Consequences and Policy Recommendations*. New York: Inter-American Development Bank.

Gambin, Lynn, Anne E. Green, and Terence Hogarth. 2009. *Exploring the Links Between Skills and Productivity: Final Report*. Coventry: Warwick Institute for Employment Research.

Gambin, Lynn, Terence Hogarth, Liz Murphy, Katie Spreadbury, Chris Warhurst, and Mark Winterbotham. 2016. *Research to Understand the Extent, Nature and Impact of Skills Mismatches in the Economy*. London: Department for Business, Innovation and Skills.

Garibaldi, Pietro. 2006. *Personnel Economics in Imperfect Labour Markets*. Oxford: Oxford University Press.

González Espitia, Carlos Giovanni, and Jhon James Mora Rodríguez. 2011. "Políticas activas de empleo para Cali-Colombia." *Estudios Gerenciales* 27, no. 118: 13–41.

González-Velosa, Carolina, and David Rosas-Shady. 2016. *Avances y retos en la formación para el trabajo en Colombia*. Bogotá: Banco Interamericano de Desarrollo.

Green, Francis. 2011. *What Is Skill?: An Inter-Disciplinary Synthesis*. London: Centre for Learning and Life Chances in Knowledge Economies and Societies.

Green, Francis, Stephen Machin, and David Wilkinson. 1998. "The Meaning and Determinants of Skills Shortages." *Oxford Bulletin of Economics and Statistics* 60, no. 2: 165–87.

Green, Francis, and Yu Zhu. 2008. "Overqualification, Job Dissatisfaction, and Increasing Dispersion in the Returns to Graduate Education." *Oxford Economic Papers* 62, no. 4: 740–63.

Grugulis, Irena, Chris Warhurst, and Ewart Keep. 2004. "What's Happening to 'Skill.'" In *The Skills That Matter*, edited by Chris Warhurst, Ewart Keep, and Irena Grugulis, 1–18. New York: Palgrave Macmillan, 2004

Guataqui, Juan Carlos, Jeisson Cárdenas, and Jaime Montaña. 2014. "La problemática del análisis laboral de demanda en Colombia." *Perfil de Coyuntura Económica* 24: 71–107.

Gweon, Hyukjun, Matthias Schonlau, Lars Kaczmirek, Michael Blohm, and Stefan Steiner. 2017. "Three Methods for Occupation Coding Based on Statistical Learning." *Journal of Official Statistics* 33, no. 1: 101–22.

Hamermesh, Daniel S. 1996. *Labor Demand*. Princeton: Princeton University Press.

Handel, Michael J. 2012. "Trends in Job Skill Demands in OECD Countries." *OECD Social, Employment and Migration Working Papers, No. 143*. Paris: OECD Publishing.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24, no. 3: 411–82.

Henson, Robin K. 2001. “Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha.” *Measurement and Evaluation in Counseling and Development* 34, no. 3: 177–89.

Holmes, David, and M. Catherine McCabe. 2002. “Improving Precision and Recall for Soundex Retrieval.” In *Proceedings. International Conference on Information Technology: Coding and Computing*, 22–26. Las Vegas: IEEE.

Huang, Anna. 2008. “Similarity Measures for Text Document Clustering.” In *Proceedings of the New Zealand Computer Science Research Student Conference* (NZCSRSC2008), 49–56. Christchurch, New Zealand.

Hughes, Deirdre, Jenny Bimrose, and Sally-Anne Barnes. n.d. “How Is Careers Labour Market Information and Intelligence Being Used and Making an Impact across the World?” Accessed July 1, 2019. http://www.lmiforall.org.uk/2017/05/how-is-careers-labour-market-information-and-intelligence-being-used-and-making-an-impact-across-the-world/.

Hussmanns, Ralf. 2004. “Statistical Definition of Informal Employment: Guidelines Endorsed by the Seventeenth International Conference of Labour Statisticians (2003).” 7th Meeting of the Expert Group on Informal Sector Statistics (Delhi Group), New Delhi, 2-4 February 2004. Accessed July 19, 2018. https://ilo.org/public/english/bureau/stat/download/papers/def.pdf.

International Labour Organization (ILO). 2003. “Guidelines Concerning a Statistical Definition of Informal Employment.” Last modified November 1, 2003. https://ilo.org/global/statistics-and-databases/standards-and-guidelines/guidelines-adopted-by-international-conferences-of-labour-statisticians/WCMS_087622/lang--en/index.htm .

____. 2008. “ISCO-08 Part 1: Introductory and Methodological Notes.” Last modified June 21, 2016. https://www.ilo.org/public/english/bureau/stat/isco/isco08/.

____. 2010. “ISCO. International Standard Classification of Occupations.” Last modified June 9, 2010. https://www.ilo.org/public/english/bureau/stat/isco/.

____. 2012. *International Standard Classification of Occupations. Structure, Group Definitions and Correspondence Tables*. Geneva: International Labour Organziation.

____. 2013. *Measurement of the Informal Economy*. Last modified March 19, 2013. https://www.ilo.org/wcmsp5/groups/public/---ed_emp/---emp_policy/documents/publication/wcms_210443.pdf.

____. 2014. *Policies for the Formalization of Micro and Small Enterprises in Colombia*. Last modified November 6, 2014. https://www.ilo.org/empent/whatsnew/WCMS_318211/lang--en/index.htm.

____. 2016a. *What Works: Active Labour Market Policies in Latin America and the Caribbean*. Accessed July 9, 2018. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_492373.pdf

____. 2016b. *2016 Labour Overview: Latin America and the Caribbean*. Lima: ILO / Regional Office for Latin America and the Caribbean.

____. 2018. "ILO Thesaurus." Last modified August 19, 2020. https://metadata.ilo.org/thesaurus.html.

____. 2019a. "ILOSTAT." https://ilostat.ilo.org/.

____. 2019b. "Issues to Be Addressed in the Revision of the Standards for Statistics on Informality." Accessed November 12, 2019. https://www.ilo.org/ilostat-files/Documents/Informality WG meeting 1 - Discussion paper.pdf.

Jones, Rini, and Peter Elias. 2004. *CASCOT: Computer-Assisted Structured Coding Tool*. Coventry: Warwick Institute for Employment Research.

Kässi, Otto, and Vili Lehdonvirta. 2018. "Online Labour Index: Measuring the Online Gig Economy for Policy and Research." *Technological Forecasting and Social Change* 137: 241–48.

Kautz, Tim, James J. Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans. 2014. *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*. NBER Working Paper No. 20749. Cambridge, MA: National Bureau of Economic Research.

Kennan, Mary Anne, Patricia Willard, Dubravka Cecez-Kecmanovic, and Concepción S. Wilson. 2008. "IS Knowledge and Skills Sought by Employers: A Content Analysis of Australian IS Early Career Online Job Advertisements." *Australasian Journal of Information Systems* 15, 2: 1–22.

Kugler, Adriana, and Maurice Kugler. 2009. "Labor Market Effects of Payroll Taxes in Developing Countries: Evidence from Colombia." *Economic Development and Cultural Change* 57, no. 2: 335–58.

Kuhn, Peter J. 2014. "The Internet as a Labor Market Matchmaker." *IZA World of Labor* 18: 1–10.

Kureková, Lucia Mýtna, Miroslav Beblavy, and Anna-Elisabeth Thum. 2014. *Using Internet Data to Analyse the Labour Market: A Methodological Enquiry*. IZA Discussion Paper No. 8555. Bonn, Germany: IZA.

Kureková, Lucia Mýtna, Miroslav Beblavy, and Anna-Elisabeth Thum. 2016. "Employers' Skill Preferences across Europe: Between Cognitive and Non-Cognitive Skills." *Journal of Education and Work* 29, no. 6: 662–87.

Laney, Doug. 2001. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Stamford: META Group.

Larsen, Christa, Sigrid Rand, Alfons Schmid, and Andrew Dean, eds. 2018. *Developing Skills in a Changing World of Work: Concepts, Measurement and Data Applied in Regional and Local Labour Market Monitoring across Europe*. München, Germany: Rainer Hampp Verlag.

Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics Doklady* 10: 707–10.

Lima, Antonio, and Hasan Bakhshi. 2018. *Classifying Occupations Using Web-Based Job Advertisements: An Application to STEM and Creative Occupations*. ESCoE Discussion Paper 2018-08. London: Economic Statistics Centre of Excellence.

Lindqvist, Erik, and Roine Vestman. 2011. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics* 3, no. 1: 101–28.

Linkedin. n.d. "Acerca de Linkedin." Accessed April 5, 2018. https://about.linkedin.com/es-es.

LinkedIn Economic Graph. 2018. "LinkedIn's 2017 US Emerging Jobs." Accessed March 29, 2018. https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report.

Little, Roderick, and Donald Rubin. 2014. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

LMI for All. n.d. "What Is Replacement Demand?" Accessed April 23, 2018. http://www.lmiforall.org.uk/2017/05/what-is-replacement-demand/.

ManpowerGroup. 2016. "Skilled Talent: It's at Your Fingertips. As Organizations Report the Highest Talent Shortage Since 2007, Employers Look to Develop Their Own Workforces to Fill In-Demand Roles." Accessed December 10, 2016. https://www.manpowergroup.com/wps/wcm/connect/8ccb11cb-1ad4-4634-84ea-1656ee74b3ed/GlobalTalentShortageSurvey-Press-Release.pdf?MOD=AJPERES&ContentCache=NONE&.

____. *The Talent Shortage*. Accessed July 19, 2018. https://web.manpowergroup.us/talent-shortage

Marinescu, Ioana, and Ronald Wolthoff. 2016. "Opening the Black Box of the Matching Function: The Power of Words." *Journal of Labor Economics* 38, no. 2: 535–68.

Maurer, Steven D., and Yuping Liu. 2007. "Developing Effective E-Recruiting Websites: Insights for Managers from Marketers." *Business Horizons* 50, no. 4: 305–14.

Mavromaras, Kostas, Josh Healy, Sue Richardson, Peter Sloane, Zhang Wei, and Rong Zhu. 2013. *A System for Monitoring Shortages and Surpluses in the Market for Skills*. Adelaide, Australia: National Institute of Labour Studies.

Mazza, Jacqueline. 2017. "Jobs and Job Search in Developing Countries: Nice Work If You Can Get It!" In *Labor Intermediation Services in Developing Economies: Adapting Employment Services for a Global Age*, 1–18. New York: Palgrave Macmillan.

McGuinness, Seamus, and Luis Ortiz. 2016. "Skill Gaps in the Workplace: Measurement, Determinants and Impacts." *Industrial Relations Journal* 47, no. 3: 253–78.

McGuinness, Seamus, and Konstantinos Pouliakas. 2016. *Deconstructing Theories of Overeducation in Europe: A Wage Decomposition Approach*. IZA Discussion Paper No. 9698. Bonn, Germany: IZA.

Ministerio de Educación Nacional, República de Colombia. 2006. *Decreto No. 1001*. Accessed July 19, 2018. https://www.mineducacion.gov.co/1621/articles-96961_archivo_pdf.pdf.

Migration Advisory Committee (MAC). 2008. *Skilled, Shortage, Sensible: The Recommended Shortage Occupation Lists for the UK and Scotland*. Accessed June 10, 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/257245/shortageoccupationlistreport.pdf.

____. 2017. *Assessing Labour Market Shortages. A Methodology Update*. Accessed May 20, 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/586110/2017_01_26_MAC_report_Assessing_Labour_Market_Shortages.pdf.

Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *Journal of Political Economy* 66, no. 4: 281–302.

Mondragón-Vélez, Camilo, Ximena Peña, and Daniel Wills. 2010. "Labor Market Rigidities and Informality in Colombia." *Economía* 11, no. 1: 65–95.

Mora, Jhon James, and Juan Muro. 2008. "Sheepskin Effects by Cohorts in Colombia." *International Journal of Manpower* 29, no. 2: 111–21.

Mortensen, Dale T. 1970. "Job Search, the Duration of Unemployment, and the Phillips Curve." *The American Economic Review* 60, no. 5: 847–62.

Mortensen, Dale T., and Christopher A. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment." *The Review of Economic Studies* 61, no. 3: 397–415.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press.

National Research Council. 2010. *A Database for a Changing Economy: Review of the Occupational Information Network (O*NET).* Washington, DC: National Academies Press.

OECD Statistics Portal. n.d. "Glossary of Statistical Terms." Accessed March 10, 2018. https://stats.oecd.org/glossary/index.htm.

Office for National Statistics (ONS). n.d. "Vacancy Survey." Last updated April 23, 2020. https://www.ons.gov.uk/surveys/informationforbusinesses/business-surveys/vacancysurvey.

____. 2015. "Labour Force Survey (LFS) QMI." Last updated January 13, 2015. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveylfsqmi.

____. 2016a. "VACS01: Vacancies and Unemployment." Last updated August 17, 2016. https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/vacanciesandunemploymentvacs01.

____. 2016b. "VACS02: Vacancies by Industry." Last updated August 17, 2016. https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/vacanciesbyindustryvacs02.

____. 2016c. "VACS03: Vacancies by Size of Business." Last updated August 17, 2016. https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/datasets/vacanciesbysizeofbusinessvacs03.

____. 2019. "Vacancies and Jobs in the UK: June 2019." Accessed June 28, 2019. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/jobsandvacanciesintheuk/june2019.

Okay-Somerville, Belgin, and Dora Scholarios. 2013. "Shades of Grey: Understanding Job Quality in Emerging Graduate Occupations." *Human Relations* 66, no. 4: 555–85.

O*NET. 2006. "New and Emerging (N&E) Occupations Methodology Development Report." Accessed September 20, 2018. http://www.onetcenter.org/reports/NewEmerging.html.

O*NET Resource Center. n.d. "About O*NET." Accessed January 24, 2019. https://www.onetcenter.org/overview.html.

Organización de Estados Iberoamericanos (OEI). n.d. "Sistemas Educativos Nacionales. Colombia." Accessed January 10, 2019. https://www.oei.es/historico/quipu/colombia/.

Organisation for Economic Co-operation and Development (OECD). 2011. *Quality Framework and Guidelines for OECD Statistical Activities*. Paris: OECD Publishing.

____. 2012. *Better Skills, Better Jobs, Better Lives: A Strategic Approach to Skills Policies*. Paris: OECD Publishing.

____. 2014a. *Education at a Glance 2014. OECD Indicators*. Paris: OECD Publishing.

____. 2014b. *Preventing Unemployment and Underemployment from Becoming Structural*. Report prepared for the G20 Labour and Employment Ministerial Meeting, Melbourne, Australia, 10-11 September 2014. Accessed August 5, 2018. https://www.oecd.org/g20/topics/employment-education-and-social-policies/OECD-Preventing-unemployment-and-underemployment-from-becoming-structural-G20.pdf.

____. 2015a. *Colombia: Policy Priorities for Inclusive Development*. Paris: OECD Publishing.

____. 2015b. *Latin American Economic Outlook 2015: Education, Skills and Innovation for Development*. Paris: OECD Publishing.

____. 2016a. *Skills Matter: Further Results from the Survey of Adult Skills*. OECD Skills Studies. Paris: OECD Publishing.

____. 2016b. *Education in Colombia: Highlights*. Accessed April 24, 2018. https://www.oecd.org/education/school/Education-in-Colombia-Highlights.pdf.

____. 2016c. *Getting Skills Right: Assessing and Anticipating Changing Skill Needs*. Paris: OECD Publishing.

____. 2016d. *OECD Reviews of Labour Market and Social Policies: Colombia*. Paris: OECD Publishing.

____. 2017a. *Financing SMEs and Entrepreneurs 2017: An OECD Scoreboard*. Paris: OECD Publishing.

____. 2017b. *Latin American Economic Outlook 2017: Youth, Skills and Entrepreneurship*. Paris: OECD Publishing.

____. 2017c. *OECD Employment Outlook 2017*. Paris: OECD Publishing.

Organización Internacional del Trabajo (OIT). 2013. *Panorama Laboral 2013 América Latina y El Caribe*. Lima: OIT / Oficina Regional para América Latina y el Caribe.

Oxford Dictionaries. 2017. "Web scraping."

Oyer, Paul, and Scott Schaefer. 2010. "Personnel Economics: Hiring and Incentives." In *Handbook of Labor Economics*, Volume 4b, edited by Orley Ashenfelter and David Card, 1769–823. Elsevier.

Özköse, Hakan, Sertaç Ari, and Cevriye Gencer. 2015. "Yesterday, Today and Tomorrow of Big Data." *Procedia-Social and Behavioral Sciences* 195: 1042–50.

Palmer, Robert. 2018. *Jobs and Skills Mismatch in the Informal Economy*. Geneva: International Labour Organization.

Perry, Guillermo E., William F. Maloney, Omar S. Arias, Pablo Fajnzylber, Andrew D. Mason, and Jaime Saavedra-Chanduvi. 2007. *Informality: Exit and exclusion*. Washington, DC: The World Bank.

Pertold-Gebicka, Barbara. 2014. "Job Market Polarization and Employment Protection in Europe." *Acta VSFS* 8, no. 2: 133–48.

Pierre, Gaëlle, Maria Laura Sanchez Puerta, Alexandria Valerio, and Tania Rajadel. 2014. *STEP Skills Measurement Surveys: Innovative Tools for Assessing Skills*. Social Protection and Labor Discussion Paper No. 1421. Washington, DC: World Bank Group.

Piore, Michael J. 1972. "Notes for a Theory of Labor Market Stratification." Cambridge, MA: MIT Press.

Psacharopoulos, George. 1985. "Returns to Education: A Further International Update and Implications." *Journal of Human Resources* 20, no. 4: 583–604.

Psacharopoulos, George. 2006. "The Value of Investment in Education: Theory, Evidence, and Policy." *Journal of Education Finance* 32, no. 2: 113–36.

Quality and Qualifications Ireland (QQI). n.d. "National Framework of Qualifications (NFQ)." Accessed June 10, 2018. https://www.qqi.ie/Articles/Pages/National-Framework-of-Qualifications-(NFQ).aspx.

Rasmussen, Karsten Boye. 2008. "General Approaches to Data Quality and Internet-Generated Data." In *The Sage Handbook of Online Research Methods*, edited by Nigel Fielding, Raymond M. Lee, and Grant Blank, 79–97. SAGE Publications Ltd.

Reich, Michael, David M. Gordon, and Richard C. Edwards. 1973. "A Theory of Labor Market Segmentation." *The American Economic Review* 63, no. 2: 359–65.

Reimsbach-Kounatze, Christian. 2015. *The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies*. 245. Paris: OECD Publishing.

Rothwell, Jonathan. 2014. *Still Searching: Job Vacancies and STEM Skills*. Washington, DC: Brookings Institution.

Rutherford, Donald. 1992. *Routledge Dictionary of Economics*. New York: Routledge.

Saavedra, Juan Esteban, and Carlos Medina. 2012. "Formación para el trabajo en Colombia." *Borradores de Economía* 740. Bogotá: Banco de la República.

Salazar-Xirinachs, José Manuel. 2017. *The Future of Work, Employment and Skills in Latin America and the Caribbean*. Last modified February 15, 2017. https://www.ilo.org/wcmsp5/groups/public/---americas/---ro-lima/---sro-port_of_spain/documents/publication/wcms_544337.pdf.

Salvatori, Andrea. 2018. "The Anatomy of Job Polarisation in the UK." *Journal for Labour Market Research* 52, no. 8.

Sánchez Molina, Eihsnover. 2013. *Clasificación nacional de ocupaciones. Versión 2013*. Bogotá: Servicio Nacional de Aprendizaje (SENA).

Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs* 6, no. 4: 317–44.

Sentz, Rob. 2013. "How Should We Look at Jobs? A Discussion of Labor Market Data and Job Postings." *Emsi Research*. Last modified April 9, 2013. https://www.economicmodelling.ca/2013/04/09/how-should-we-look-at-jobs-a-discussion-of-labor-market-data-and-job-postings/.

Servicio Nacional de Aprendizaje (SENA). 2015. *Informe: Operación estadística para seguimiento a las condiciones de empleabilidad y desempeño de los egresados del SENA*. Bogotá: SENA.

Shah, Chandra, and Gerald Burke. 2003. *Skills Shortages: Concepts, Measurement and Implications*. Melbourne, Victoria: Centre for the Economics of Education and Training (CEET).

Smith, Aaron. 2015. "Searching for Work in the Digital Era." *Pew Research Center Internet & Technology*. Last modified November 19, 2015. https://www.pewresearch.org/internet/2015/11/19/searching-for-work-in-the-digital-era/

Spence, Michael. 1978. "Job Market Signaling." In *Uncertainty in Economics. Readings and Exercises,* edited by Peter Diamond and Michael Rothschild, 281–306. Burlington: Elsevier Science.

Spitz-Oener, Alexandra. 2006. "Technical Change, Job Tasks, and Rising Educational Demands: Looking Outside the Wage Structure." *Journal of Labor Economics* 24, no. 2: 235–70.

State, Bogdan, Mario Rodriguez, Dirk Helbing, and Emilio Zagheni. 2014. "Migration of Professionals to the U.S. Evidence from LinkedIn Data." In *Social Informatics. SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers*, edited by Luca Maria Aiello and Daniel McFarland, 531–43. Geneva: Springer International.

Štefánik, Miroslav. 2012. "Internet Job Search Data as a Possible Source of Information on Skills Demand (with Results for Slovak University Graduates)." In *Building on Skills Forecasts—Comparing Methods and Applications. Conference Proceedings*, edited by Cedefop, 246–60. Luxembourg: Publications Office of the European Union.

Stiglitz, Joseph E., Carl E. Walsh, Jeffrey Gow, Ross Guest, William Richmond, and Max Tani. 2013. *Principles of Economics*. Melbourne: Wiley Australia.

Stopher, Peter. 2012. *Collecting, Managing, and Assessing Data Using Sample Surveys*. Cambridge: Cambridge University Press.

Streiner, David L. 2003. "Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency." *Journal of Personality Assessment* 80, no. 1: 99–103.

Swier, Nigel. 2016. "Web Scraping for Job Vacancy Statistics." Paper presented at the ESS Big Data Workshop 2016, Ljubljana, Slovenia, October 13-14, 2016. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/d/d4/BDES_2018_WP1_Presentation.pdf.

Tijdens, Kea, Miroslav Beblavy, and Anna Thum-Thysen. 2015. *Do Educational Requirements in Vacancies Match the Educational Attainments of Job-Holders? An Analysis of Web-Based Data for 279 Occupations in the Czech Republic*. InGRID working paper; No. MS 21.7. Leuven: Research Institute for Work and Society, KU Leuven.

Turrell, Arthur, Bradley Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. 2018. *Using Job Vacancies to Understand the Effects of Labour Market Mismatch on UK Output and Productivity*. Staff Working Paper No. 737. London: Bank of England.

Turrell, Arthur, Bradley J. Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. 2019. *Transforming Naturally Occurring Text Data into Economic Statistics: The Case of Online Job Vacancy Postings*. NBER Working Paper No. 25837. Cambridge, MA: The National Bureau of Economic Research.

UK Commission for Employment and Skills (UKCES). 2012. *Developing Occupational Skills Profiles for the UK: A Feasibility Study*. Evidence Report 44. Last modified February 2012. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/306764/ER44_Developing_occupational_skills_profiles_for_the_uk_a_feasibility_study_-_Feb_2012.pdf.

____. 2014. *The Future of Work: Jobs and Skills in 2030*. Evidence Report 84. Last modified February 2014. https://www.oitcinterfor.org/sites/default/files/file_publicacion/thefutureofwork.pdf.

____. 2016. *Employer Skills Survey 2015: UK Results*. Evidence Report 98. Last modified May 2016. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/525450/UKESS_2015_Technical_Report__May_.pdf.

UK Data Service. n.d. "Average Duration of Unemployment (2018 Edition)." Accessed May 12, 2019. https://stats2.digitalresources.jisc.ac.uk/Index.aspx?DataSetCode=AVD_DUR.

Valencia, Ferney Hernando, Carlos Alberto Suárez Medina, Carlos Rocha Ruíz, and Dora Alicia Mora Pérez. 2016. "Composición de la economía de Bogotá." *Revista del Banco de la República* 89, no. 1069: 11–36.

Vallas, Steven Peter. 1990. "The Concept of Skill: A Critical Review." *Work and Occupations* 17, no. 4: 379–98.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28, no. 2: 3–28.

Verougstraete, Remie. 2018. "NEW: Skills Taxonomy Update." *Emsi Research*. Last modified June 14, 2018. https://www.economicmodeling.com/2018/06/14/new-skills-taxonomy-update/.

Wageindicator.org. 2009. *EurOccupations: CASCOT Software for Coding Job Titles*. European Policy Brief No. 3. Accessed May 25, 2018. https://wageindicator.org/copy_of_documents/policy-briefs/European-Policy-Brief-no-3-CASCOT-coding-program-EUROCCUPATIONS-20100104.pdf.

Warhurst, Chris, Ken Mayhew, David Finegold, and John Buchanan. 2017. *The Oxford Handbook of Skills and Training*. Oxford: Oxford University Press.

Warwick Institute for Employment Research (IER). 2018. "Cascot International." Accessed September 20, 2018. https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/.

Williams, Richard D. 2004. "The Demand for Labour in the UK An Introduction to the Topic Illustrated with Data from Two Regions." *Labour Market Trends* 112, no. 8: 321–30.

World Bank. 2010. *Informality in Colombia. Implications for Worker Welfare and Firm Productivity*. Washington, DC: The World Bank.

____. n.d. "Education Statistics: Education Attainment." Accessed September 18, 2018. http://databank.worldbank.org/data/reports.aspx?source=Education-Statistics:-Education-Attainment.

World Bank, International Comparison Program. 2019. "GDP per capita, PPP (current international $) - Colombia, Germany, OECD members." Accessed January 20, 2019. https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?locations=CO-DE-OE.

Zhang, Cha, and Yunqian Ma. 2012. *Ensemble Machine Learning: Methods and Applications*. New York: Springer-Verlag.



Appendix

Appendix A: Examples of Job Portal Structures

Figure A.1 shows how differently two websites ("Jobportal_a" and "Jobportal_c") present their vacancies. The boxes in panel A and B highlight the job vacancy attributes that each website displays in listed job advertisements.

There are some things in common between these two job portals. In the Jobportal_a and Jobportal_c panels, Box A highlights the job title, which is a short description about the position to be filled. However, there are also some differences between each website. Jobportal_a displays the name of the company that advertises the job, and the city where the vacancy (or vacancies) is available in Box B. In Box C, a brief description of the job vacancy (e.g. level of education required by the employer) is shown, and Box D displays when the job was advertised. In contrast, Jobportal_c displays the job title in Box A (as mentioned above), the department/city where the vacancy is available (Box B), and the date when the vacancy is going to expire (Box C). There is no Box D on this job portal.

Figure A.1. **Job portal comparison**[1]

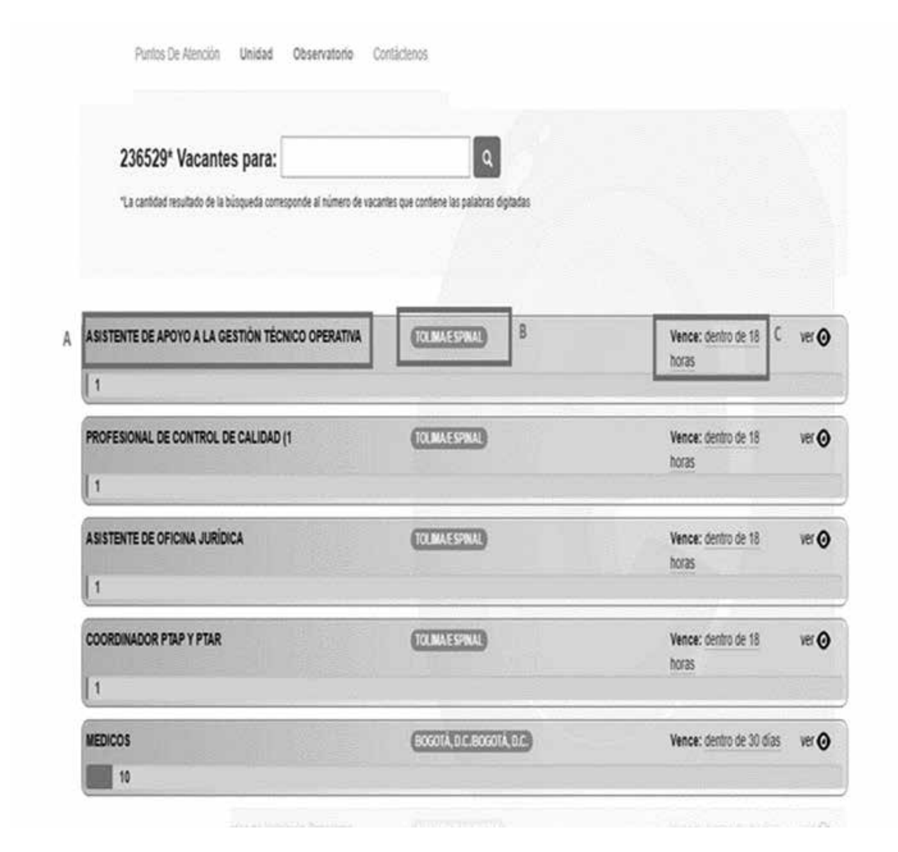
Panel A: Jobportal_a[2]



1 The figures are presented in Spanish, because this is the original language used on Colombian job boards.

2 Translation of relevant parts of the text: Box A: Sales assistants; Box B: Company's name: M&R Selectiva, City: Bogotá; Box C: A recognized automotive diagnosis centre requires bachelors, technicians or technologists in administrative areas; Box D: Today, at 06:11 a.m.

Panel B: Jobportal_c[3]



Sources: Jobportal_a and Jobportal_c.

Moreover, when information about each job is examined in more detail, greater differences can be observed regarding how the information is presented between and within job portals. As seen in Figure A.2, information on the same website (in this case Jobportal_a) might vary from one advertisement to another. Panel A displays a job vacancy for a "Computer, automated teller, and office machine repairer," while Panel B requires a "Labourer and freight, stock, and material mover." Panel A shows information about the job title, experience required, wage offered, city and department (where the vacancy is available), and the date when the advertisement was published. In comparison,

3 Translation of relevant parts of the text: Box A: Assistant management operator; Box B: Department: Tolima, City: Espinal; Box C: Expiration date: in 18 hours.

Box A in Panel B displays the job title, city and department (where the vacancy is available), and date when the advertisement was published. Consequently, information about required experience is not shown in Box A. However, information regarding experience for this vacancy can be found in Box C (at the bottom of the website).

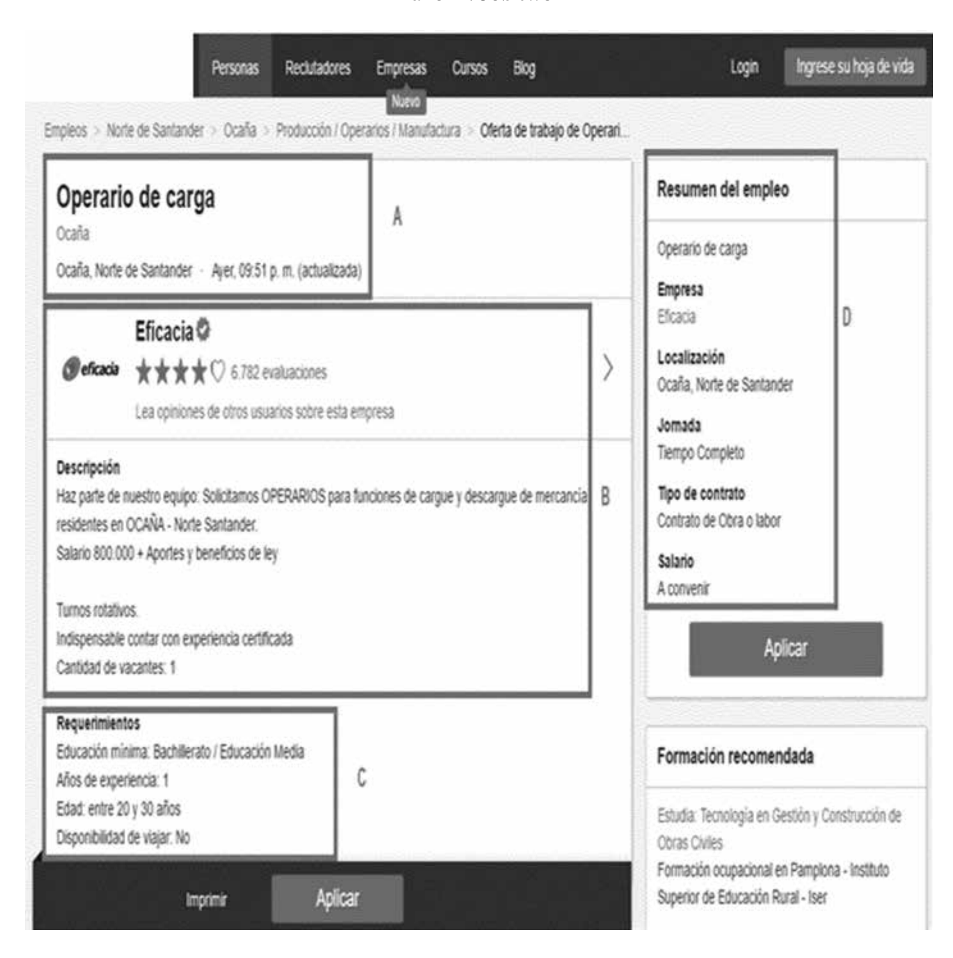
Figure A.2. **Job advertisement comparison within the same job portal**

Panel A: Job one[4]



4 Box A includes the following information: System support technicians. Minimum six months of work experience. Wage 782,000 Colombian pesos (monthly). City: Medellín. Department: Antioquia. Posted: yesterday at 09:51 p.m.; Box B: Important company in the service sector. Description: System support technicians or similar are required with a minimum of six months of work experience to repair printers and photocopiers. The tasks are assembly and maintenance of equipment through business visits. Date of hire: 30/06/2018. Number of jobs: 4. Box C: Requirements. Minimum bachelor certificate required. No travel is required. Box D: Job summary. System support technicians. Localisation: Medellín, Antioquia. Working day: Full-time. Type of contract: Indefinite term contract. Wage: 782,000 Colombian pesos (monthly).

Panel B: Job two[5]



Source: Jobportal_a.

[5] Box A includes the following information: Cargo operator. City: Ocaña. Department: Norte de Santander. Posted: yesterday at 09:51 p.m.; Box B: Company's name: Eficacia. Description: Cargo operators are required to load and unload merchandise. Wage: 800,000 Colombian pesos (monthly). Rotating work shifts. Certified experience is required. Number of jobs: 1. Box C: Requirements. Minimum bachelor certificate required. Years of experience: 1. Age: Between 20 and 30 years old. No travel is required. Box D: Job summary. Cargo operator. Company's name: Eficacia. Localisation: Ocaña, Norte de Santander. Working day: Full-time. Type of contract: Task type contract. Wage: To be agreed upon.

Panel A in Figure A.3 shows the HTML code of a Jobportal_a job advertisement. Information about the job title and company are delimited by the tags </p> and <h3>, respectively. In Panel B, the HTML code is displayed for a job advertisement on “jobportal_c.gov.co.” On this website, the job title is defined by the syntax “<h2><span id=’ctl00_ContenidoPanel_ucdetalle_oferta_lblTituloCargo’>” and company information by the syntax <h4><span id=”ctl00_ContenidoPanel_ucdetalle_oferta_lblEmpresa”>Empresa:.

Figure A.3. **Code comparison between job portals**

Panel A: Jobportal_a

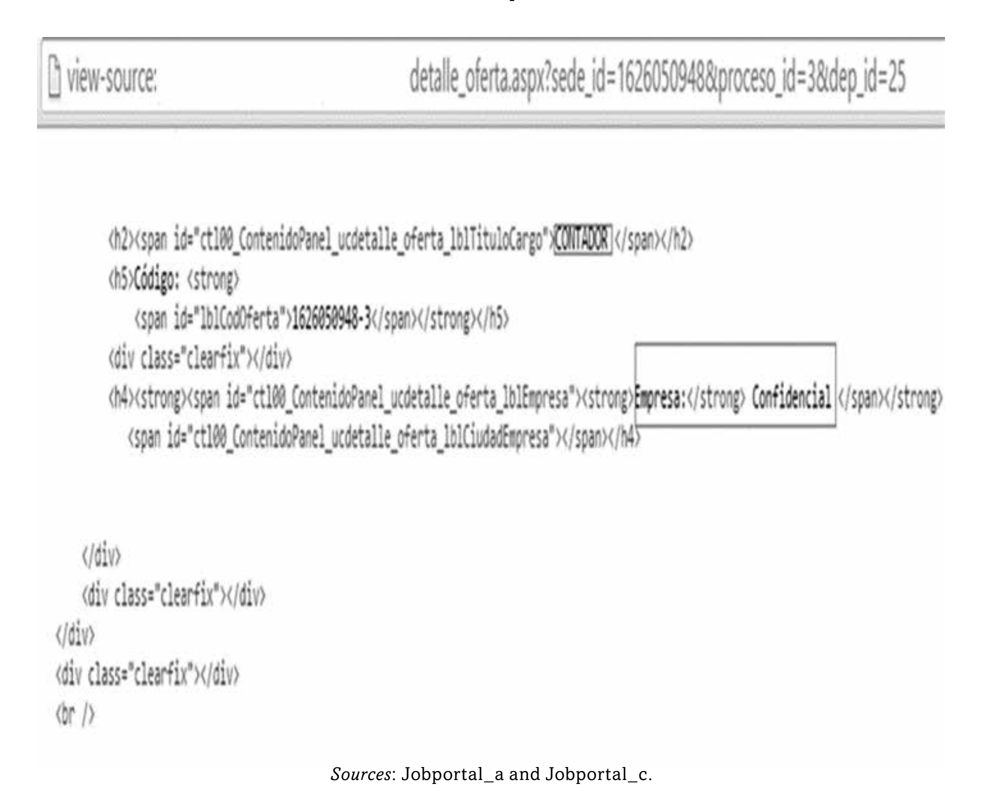
```
view-source:https                    ofertas-de-trabajo/

    <div class="box_r_titulo">
        <p>Resumen del empleo</p>
    </div>

<ul>
    <li>
        <p>
            Contador</p>
    </li>

        <li>
            <h3>Empresa</h3>
            <p>
```

Panel B: Jobportal_c



Sources: Jobportal_a and Jobportal_c.

For illustrative purposes, Figure A.4 shows the HTML code structure of the web page Jobportal_b for a job advertisement. The web scraping technique recognises tags (or Xpath) where relevant information exists. In Box A, the tag "class='js-jobOffer-title'" contains the information related to a job title ("*Contador Senior*"); in Box B, the tag "class=='js-jobOffer-salary'" ("*$2,5 a $3 millones*") describes the salary offered; and the tags "itemprop='addressCountry'" and itemprop='addressRegion'" in Boxes C and D, respectively, contain information regarding the country and the region where the vacancy is offered (Colombia and Bogotá, respectively[6]). Consequently, the R codes (developed in this book) recognise each one of those tags (or Xpath) and extract the information of interest for each job advertisement from each job portal.

---

6 As shown in Chapter 7, as expected, most of the vacancies are available in Colombia. However, there are some offers to work in other countries.

Figure A.4. **HTML code structure**

```
view-source:          /co/ofertas-trabajo/contador-senior/1883464899

<h2 itemprop="title" class="ee-offer-detail-modal-title js-offer-title">
    <span class="js-jobOffer-title" data-take-keyword="CONTADOR SENIOR ">    A
        Contador senior
    </span>
</h2>

<div class="row">
    <div class="col-xs-12 col-sm-6 data-column hidden-xs ee-quick-offer-data">
        <p>
            <i class="fa fa-usd fa-fw"></i>
                <span itemprop="baseSalary" itemscope itemtype="http://schema.org/MonetaryAmount">    B
                    <span itemprop="value" class="js-joboffer-salary">
                        $2,5 a $3 millones
                    </span>
                    <span class="hide" itemprop="currency">
                        COP
                    </span>
                </span>
        </p>

        <p>
            <i class="fa fa-map-marker fa-fw"></i>
            <span itemprop="jobLocation" itemscope itemtype="http://schema.org/Place">
                <span>
                        <span itemprop="address" class="hidden" itemscope itemtype="http://schema.org/PostalAddress">
                            <span itemprop="addressCountry">    C
                                Colombia
                            </span>
                            <span itemprop="addressRegion">    D
                                Bogot&#225; D.C.
                            </span>
```

Source: Jobportal_b.

It is important to note that sometimes companies do not advertise information regarding a characteristic of a vacancy such as salary using the corresponding tag (e.g. “class==‘js-jobOffer-salary’”). In these cases, the algorithm searches for the tag, and when the tag is not found, the algorithm leaves a missing value in the database. This issue does not necessarily mean that there is no information regarding a certain job characteristic (e.g. salary). Employers might have posted the job characteristics using other tags, for instance, in the job description tag (“class==‘offer-detail’”). Indeed, it is common to see that employers post most of the relevant information using the tag description of the vacancy (which is a paragraph where companies describe the job position), while other tags, such as for salary, might not be used. Consequently, as will be seen in Appendix B, the implementation of text mining techniques is necessary to identify all the relevant information in job advertisements.

Appendix B: Text Mining

Columns 1 and 2 in Table B.1 show part of the information provided by employers for a pair of job vacancies. In the job description (Column 1) for the first vacancy, the employer indicates that they require a person with an undergraduate certificate; however, in the website column, where the employer was supposed to provide specific information regarding educational requirements (Column 2), there is a missing value. In contrast, for the second vacancy (see the second row of Table B.1), in the job description (Column 1), the employer does not specify any educational requirement. However, for this vacancy, information regarding education is available in the "Educational requirements" column (Column 2).

Thus, the algorithm needs to "read" different columns in the scraped data (not only the educational requirements column) to identify qualification requirements. Some of the relevant information might be only mentioned in the job description or in other specific columns; additionally, information might be repeated in different columns. In this example, the algorithm creates a "Dummy_Undergraduate_certificate" and a "Dummy_PhD_certificate" column (third and fourth columns in Table B.1). Based on the information provided in the job description, the "Dummy_Undergraduate_certificate" column identifies (by taking a value of one) that the first vacancy requires a person with an undergraduate certificate, while the second vacancy does not require a person with this certificate. On the other hand, based on the "Educational requirements" column, the "Dummy_PhD_certificate" column identifies that the second vacancy (second row of Table B.1) requests a person with a PhD. It is important to note that employers might be indifferent about educational levels or another job characteristic. For instance, a vacancy might require a person with a high school or undergraduate level. In these cases, the dummy variables ("High school" and "Undergraduate_certificate") take the value of one at the same time.

Table B.1. **Example of the content of a scraped database**

Job description	Educational requirements	Dummy_ Undergraduate_ certificate	Dummy_ PhD_certificate
• Must have some previous production experience in a book publishing or printing environment • Excellent communication and interpersonal skills and consultative customer care approach • **Undergraduate certificate** • Strong IT skills with experience of using a CRM system advantageous	No information provided by the employer (missing value)	1	0
• Candidates with expertise in cancer genomics and related disciplines, and candidates with experience of working with clinical data are particularly encouraged to apply. • Editorial experience is not required, although applicants with significant editorial experience are encouraged to apply and will potentially be considered for a Senior Editor position.	A PhD (or equivalent) in a field related to cancer and/or genomics and significant research experience	0	1

Source: Author's calculations.

Additionally, there is an issue when looking for patterns in Spanish language. In this language, nouns have a gender; for instance, an undergraduate might be called "*Universitario*" (for men) or "*Universitaria*" (for women). Given this fact, and the usage of synonyms to express a job requirement, the algorithm looks for patterns in the root of the words.[7] The selection of the root of the words is a critical process where it is necessary to carefully choose the proper root to correctly classify job requirements. In this way, the dummy variables created are guaranteed to correctly identify employer requirements, even in the presence of synonyms, gendered nouns, etc.

[7] For instance, in the case of undergraduate, the algorithm looks for "universi" among other word roots.

Appendix C: Detailed Process Description for the Classification of Companies

C.1. Manual coding

Manual coding is a process through which a person (or a group of people) manually assign a code (or category) to each observation based on data characteristics. Similar to the coding of job titles (see Chapter 6), the full manual coding of each company sector is a time-consuming task. There are a few companies (most of them related to office administration, office support, and other business support activities) that post relatively more vacancies than others. Consequently, a manual coding process was conducted to ensure that the most frequently used versions of company names were properly classified. Fifty company names were coded manually.[8] These companies represent around 13% of the total number of companies listed in the vacancy database. Once this process was completed, the next step was the implementation of automatic coding using word-based matching methods.

C.2. Word-based matching methods ("Fuzzy merge")

Different word-based matching methods exist. The differences between one method and another are due to the rules employed to obtain a similarity score. For instance, the Levenshtein distance algorithm (Levenshtein 1966) calculates the distance (similarity) between two words or sentences (strings) based on the minimum number of characters that are necessary to change one word (or phrase) into other word (or sentence). As an illustration, when the Levenshtein distance algorithm compares two words such as "butcher" and "butchery," the distance will be one, which is the number of characters required to transform the word "butcher" into "butchery." Moreover, other algorithms are also

8 In the case of multi-product/sector companies, the Single Business Registry (RUES for its acronym in Spanish) only registers the main activity reported by the company. Consequently, the ISCO code assigned to those companies corresponds to the main economic activity reported in the RUES.

available, such as Soundex, in which metrics are based on the sound of the words rather than the characters of the words (see Holmes and McCabe 2002 for a detailed explanation of the Soundex algorithm).

Given the numerous algorithms available, this book used the following steps to merge company names for each job with the information available in the RUES database. First, observations that shared the same names in the vacancy and the RUES database were merged. Only 2% of the companies in the vacancy database were matchable with their ISIC code. As mentioned above, this low merge rate might be due to differences between company names in the two databases. Consequently, it was necessary to use fuzzy merge algorithms to correctly assign an ISIC code to the maximum number of company names efficiently. More specifically, the Jaro-Winkler algorithm was used to merge the databases (see Cohen, Ravikumar, and Fienberg 2003 for a detailed explanation of the Jaro-Winkler algorithm). A threshold of 98%[9] of similarity was selected. With this method, 15% of the vacancy database received an ISIC code.

For the remaining database, a Cosine similarity algorithm was implemented at a 98% matching threshold (see Huang 2008 and Appendix G for a detailed explanation of the Cosine similarity algorithm). By doing so, a further 9% of the observations in the vacancy database were assigned an ISIC code. Additionally, the Soundex algorithm at a 98% matching threshold was executed. With this procedure, 6% of the observations in the vacancy database were assigned an ISIC code.

Finally, the Levenshtein algorithm was implemented. Given the characteristic of this algorithm of being sensitive to the length of words, it was executed only for those observations that had not been merged and for which the length of a company name (number of characters) was more than 5. With this algorithm, 3% of the job announcements were merged with the RUES database. Therefore, 48% of the observations in the vacancy database (manually and with word-based matching methods) received an ISIC code, up to this point (see Figure C.1 for a summary). As can be observed, despite these word-based matching methods, there is still a considerable number of observations without an ISIC code (52%). This outcome might be due to important

9 It is important to note that matching thresholds were relatively high in order to guarantee an acceptable accuracy level.

differences between company names in the vacancy and the RUES database. Or the issue might reside in the RUES database, where a considerable number of companies might not be registered.

Figure C.1. **Fuzzy merge: The classification of companies**

Source: Author's calculations.

C.3. A return to manual coding

As a considerable number of observations (52%) were not coded with word-based matching methods, it was necessary to return to a manual coding process. As in Subsection C.1, first, coding was carried out via a visual inspection of uncoded information in the vacancy database, e.g. companies that were not coded by using the word-based matching and manual coding methods mentioned in the previous subsection. Subsequently, those company names that appeared more frequently in the database were manually coded: a total of 50. Yet, these companies represent only 4% of the vacancy database; therefore, even with the above procedure there is a considerable number of observations without an ISIC code (48%).

Thus, and as the last step, company names were used to assign ISIC codes. Frequently, company names reveal the sector where they perform their activities. This is the case, for instance, of restaurants (McDonald's restaurant) and universities (University of Santander), among others. Consequently, by using keywords from company names, it is possible to assign a code to a considerable number of job announcements. With this last procedure, it was possible to assign an ISIC code to 9% of the vacancy announcements.

Appendix D: Machine Learning Algorithms

To assign occupational codes, the basic machine learning model starts by transforming job titles into numeric values. A variable takes the value of one if the word j occurs in document i. By applying the above transformation to all x documents in a database, the result is a word co-occurrence matrix that indicates which words appear in each document. For instance, as shown in Table D.1, one job title might be "Web designer," while another would be "Network designer." Thus, three indicating variables ("Web," "Network," and "Designer") are created that will take values of one if the words appear in the job title, and zero otherwise.

Table D.1. **N-grams based on job titles**

ISIC code	Job title	Web	Designer	Network
2166	Web designer	1	1	0
2523	Network designer	0	1	1

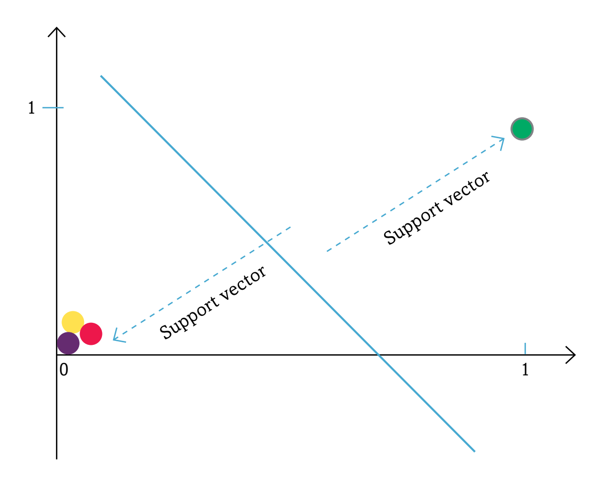
Source: Author's calculations.

This "bag of words" representation is a key element for the algorithm to learn how to classify job titles into occupations. Moreover, the computer requires an algorithm to classify job titles into occupational codes based on this bag of words. There are different kinds of algorithms—such as linear and logistic regression, random forest, naïve Bayes, among others—that help to map x inputs into the outputs y (where). The choice between one algorithm or another depends on the problem to be solved and the available data.

Appendix E: Support Vector Machine (SVM)

SVM is a machine learning algorithm. Briefly, based on the inputs (*x*), the SVM algorithm tries to find a hyperplane (or hyperplanes) (which, for simplicity, can be thought of as a line in a Cartesian plane) that best divides the data into two or more classes. In the case of job titles, the SVM algorithm considers the numeric transformation of job titles (x_{ij} variables) to create the hyperplane(s) that separate(s) the data in different occupational groups. For instance, Figure E.1 shows a simple illustration of how SVM works to classify a job title such as "Gardener" into an occupational code. In this example, a vacancy database has an observation listed with the job title "Gardener" (occupational code 6113: Gardeners, horticultural, and nursery growers), while the job titles in other observations are "Accountant" (4311: Accounting and bookkeeping clerks), "Economist" (2631: Economists), and "Psychologists" (2634: Psychologists) (indicated by the purple, yellow, and red points, respectively, in Figure E.1). For simplicity, x input (x-axis) is the numeric transformation of the word "gardener" (which takes the value of one if the word "gardener" occurs in vacancy advertisement i). Moreover, the outcome *y* takes the value of one if the occupational group is "6113: Gardeners, horticultural, and nursery growers" (green point in Figure E.1).

Figure E.1. **SVM classification with job titles**



Source: Author's calculations.

In this example, it is relatively easy to find different hyperplanes that have divided the dataset into two parts (numbers of classes). However, the best hyperplane that separates the data is the one that maximises margins between the points (0,0) and (1,1) on the Cartesian plane. These two points are named the support vectors, which are the points nearest to the hyperplane. Consequently, the greater the distance between the margins and the hyperplane, the higher the probability of correct classification. Thus, in this way, the computer learns how to classify job titles into occupations. Clearly, the algorithm needs to do more complex tasks when data contain a considerable number of observations, classes, and explanatory variables (x).

Appendix F: SVM Using Job Titles

As mentioned above, the basic machine learning approach uses job title information to assign occupational codes. This section evaluates the possibility of predicting the remaining job titles of the Colombian database by conducting an underlying machine learning approach using the SVM method. The job titles categorised in Subsections 6.4.1 to 6.4.6 were used to train and test the algorithm. Additionally, the cleaned job titles (Subsection 6.4.2) were transformed into a word co-occurrence matrix with around 4,000 terms. As mentioned above, the algorithm uses this bag of words as an input (x) to classify job titles into occupations. As observed via a manual check, 40% of the job titles were incorrectly classified. Therefore, the SVM machine learning algorithm, which only uses job titles, is not an option to classify the remaining observations in the vacancy database.

Appendix G: Nearest Neighbour Algorithm Using Job Titles

There are different ways to measure the distance between two strings (see Appendix C); however, as Gweon et al. (2017) note, one of the most common approaches is the Cosine similarity. This approach takes the vector representation of two documents (e.g. a and b) and calculates the distance between vectors in the following way:

$$Similarity(A,B) = \frac{A*B}{||A||*||B||} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

In the above formula, A and B are the vector representations of documents a and b. This similarity score can lie between 1 and 0. The greater the similarity between the two documents, the closer the value is to 1. For instance, Tables G.1 and G.2 are examples of a training dataset with a new observation to be coded. In both tables, the new record to be coded is "Web designer."

The training database in Table G.1 has a unique observation in which the job title is "Web designer" and the ISCO code 2166 ("Graphic and multimedia designers") has been assigned to this title. Consequently, the vector representation of the training database and the new record are shown in columns 4 and 5 of Table G.1.

Table G.1. **Vector representation, example one**

Source	ISCO code	Job title	Web	Designer
Training dataset	2166	Web designer	1	1
New record	…	Web designer	1	1

Source: Author's calculations.

Consequently, the similarity score between the two records for "Web designer" in Table G.1 is given by the following formula:

$$Similarity\ (new\ vacancy, row1) = \frac{(1*1) + (1*1)}{\sqrt{1^2 + 1^2} * \sqrt{1^2 + 1^2}} = 1 \quad (2)$$

In contrast, the training database in Table G.2 is composed of a unique observation in which the job title is "Network designer," and the ISCO code assigned to this title is 2523. Consequently, the vector representation of the training database and the new record are shown in columns 4 and 5 of Table G.2.

Table G.2. **Vector representation, example two**

Source	ISCO code	Job title	Web	Designer	Network
Training dataset	2523	Network designer	0	1	1
New record	...	Web designer	1	1	0

Source: Author's calculations.

The similarity score between the two records is given by:

$$Similarity\ (new\ vacancy, row2) = \frac{(0*1) + (1*1) + (1*0)}{\sqrt{1^2 + 1^2} * \sqrt{1^2 + 1^2}} = 0.5 \quad (3)$$

Therefore (as expected), the most similar observation in the training database for a new entry with a job title of "Web designer" is the one with the same string values. The new record would receive the occupational code 2166 ("Graphic and multimedia designers"). As Gweon et al. (2017) argue, the reason for the higher accuracy of the nearest neighbour algorithms compared to the SVM algorithms is that the former are more effective when the accuracy of the result highly depends on local points (very similar job titles), and little information can be obtained from remote records (dissimilar job titles).

In this regard, Gweon et al. (2017) propose an improvement using the Cosine similarity score in the following way: they denote a new document (job title) as x, the number of the nearest neighbours in the training dataset as *K(X)*, and *s(x)* the (Cosine) similarity score of the nearest neighbours (job titles). Additionally, $ki(x)$ out of the *K(X)* neighbours have the class (occupational code) c_i (i=1,2,3,…L). Consequently, the rule to assign an occupational code is defined as:

$$\gamma(c_i | x) = p(c_i | x)s(x) = \left(\frac{K(X)}{K(X) + 0.1}\right) \quad (4)$$

As Gweon et al. (2017) note, the predicted code only depends on $p(c_i|x)$. The terms $K(X)$ and $s(x)$ are constant for any observation in the training database with one string in common with the new record. Both $K(X)$ and $s(x)$ terms help to identify which new records have enough and similar neighbours to make a proper comparison. The multiplier $s(x)$ indicates the degree of closeness, a high value of $s(x)$ indicates that the job titles in the training database are very similar to the new record(s). Moreover, the term $\left(\frac{K(X)}{K(X)+0.1}\right)$ is a control for the number of neighbours. The higher this indicator, the larger the number of nearest neighbours for the new record(s). Consequently, it is supposed that the algorithm's output will be more precise when there are more nearest neighbours ($K(X)$). As Gweon et al. (2017) mention, at most, this multiplier can reduce the score by about 10% (when $K(X) = 1$), while $p(c_i|x)$ and $s(x)$ can lower γ to zero.

Consequently, for this algorithm, the term $\left(\frac{K(X)}{K(X)+0.1}\right)$ has relatively less importance than the other terms. The constant 0.1 serves to decrease the importance of this term in the total score γ. The choice of 0.1 might be seen as arbitrary. However, a smaller constant makes the γ score more sensitive to changes in $K(X)$, which are not desirable due to the other two terms ($s(x)$ and $p(c_i|x)$), which are relatively more important for this algorithm. On the other hand, a larger constant makes the γ score less sensitive to changes in $K(X)$, which would make irrelevant the term $\left(\frac{K(X)}{K(X)+0.1}\right)$.

Finally, the class in which the score has the highest values will be assigned to the new record. Table G.3 illustrates how this algorithm works for a given training database. There is a new record with the following words in the job title "Technician assistant food." There are four other observations in the training database that have one word in common with the new record. Columns 2, 3, and 4 show the vector representation of the training database and the new record. Three-quarters of the observations in the training database coincide with the new record due to the word "food." These observations have the occupational code 9412 ("Kitchen helpers"). Additionally, another observation in the training data set coincides with the new record due to the word "assistant" and has the occupational code 5223 referring to "Shop sales assistants." Following Equation 4, the values for $p(c_{9412}|x)$ and for $p(c_{5223}|x)$ are 0.75 (3/4) and 0.25, respectively. Thus, the observations with the occupational code 9412 receive the same $p(c_i|x)$ value (0.75), while the observation with the occupational code 5223 receives the value 0.25 (Column 6 of Table G.3).

For each observation in the training database, the $s(x)$ (Cosine similarity) with the new record is given by $\frac{1}{\sqrt{3}\sqrt{3}} = 0.5774$ (Column 7). Finally, the term $\left(\frac{K(X)}{K(X)+0.1}\right) = \left(\frac{4}{4+0.1}\right)$ is equal to 0.9756 for each observation (Column 8). By multiplying the terms $\gamma(c_{9412}|x)$ is equal to 0.4225, while the $\gamma(c_{5223}|x)$ score is equal to 0.1408 (Column 9). Thus, the occupational code assigned to the new record "Technician assistant food" is 9412 ("Kitchen helpers") (Column 5 of Table G.3).

Table G.3. **Nearest neighbour algorithm (Gweon et al. 2017)**

Source	Technician	Assistant	Food	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	9412	0.75	0.5774	0.9756	0.4225
	0	0	1					
	0	0	1					
	0	1	0	5223	0.25			0.1408
New record	1	1	1	9412*				

Source: Author's calculations.

* Final occupational code assigned to the new record.

However, when the training database is unbalanced, namely, there are some job titles that are considerably more common than others (for instance, shop assistant for the Colombian case), this algorithm might fail. Table G.4 shows an example of this issue. Supposing that there is a training database with four observations, the job title of one of those observations is "Help preparation food," and the job title of the remaining observations in the training database is "Shop sales assistant." Additionally, there is a new record with the following words in the job title "Technician assistant food." Columns 2 to 8 of Table G.4 show the vector representation of the training database and the new record. Three-quarters of the observations in the training database coincide with the new record due to the word "assistant." These observations have the occupational code 5223 ("Shop sales assistants"). Additionally, another observation in the training data set coincides with the new record due to the word "food" and has the occupational ISCO code 9412 for "Kitchen helpers" (Column 9). Following Equation 4, the values for $p(c_{9412}|x)$ and for $p(c_{5223}|x)$ are 0.25 (1/4) and 0.75, respectively (Column 10).

For each observation in the training database, the $s(x)$ (Cosine similarity) with the new record is given by $\frac{1}{\sqrt{3}\sqrt{3}}$ (Column 11). Finally, the term $\left(\frac{K(X)}{K(X)+0.1}\right)=\left(\frac{4}{4+0.1}\right)$ is equal to 0.9756 for each observation (Column 12). By multiplying the terms, $\gamma(c_{9412}\,|x)$ is equal to 0.0812, while the $\gamma(c_{5223}\,|x)$ score is equal to 0.2438 (Column 13). Thus, the occupational code assigned to the new record "Shop sales assistant" is 5223 (Column 9 of Table G.4). Consequently, when the training database is unbalanced (there are many job titles with the same word), the algorithm might classify all the new records that contain the word "assistant" in the category of "Shop sales assistants" (ISCO code 5223). Consequently, the accuracy level of the algorithm decreases. There are observations such as "Technician assistant food" that do not receive the proper occupational code.

One possibility to avoid this issue is to perform a resampling of the training database to balance occupational groups. However, this approach might not be effective either. Even when two observations with the word "assistant" are dropped, the predicted result for the new record will be the same as before ("Shop sales assistants," ISCO 5223). Additionally, this re-balancing affects the occupational structure of the training database and some job titles that before were correctly predicted. With this adjustment, some of them might be misclassified.

Tables G.5 and G.6 illustrate an example of this method. For the new record "Technician assistant food," there are six observations in the training database that have one word in common in the job title. With this information, the algorithm proposed in Gweon et al. (2017) would incorrectly code the new record in the "Shop sales assistants" (5223) category, as shown in Table G.5.

Table G.4. **Limitation of the nearest neighbour algorithm**

Source	Technician	Assistant	Food	Preparation	Help	Sales	Shop	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	1	1	0	0	9412	0.25	0.333	0.9756	0.0812
	0	1	0	0	0	1	1	5223	0.75			0.2438
	0	1	0	0	0	1	1					
	0	1	0	0	0	1	1					
New record	1	1	1	0	0	0		5223*				

Source: Author's calculations.

Table G.5. **An extension of the nearest neighbour algorithm (Part 1)**

Source	Job title			Job description (skills)			Parameters				
	Technician	Assistant	Food	Dispose waste	Use food cutting tools	Clean surfaces	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\gamma(c_i\|x)$
Training dataset	0	0	1	0	1	1	9412	0.333	0.5774	0.983	0.189
	0	0	1	1	0	1	9412		0.5774	0.983	0.189
	0	1	0	1	0	0	5223	0.5	0.5774	0.983	0.283
	0	1	0	0	0	0	5223		0.5774	0.983	0.283
	0	1	0	0	0	0	5223		0.5774	0.983	0.283
	1	0	0	0	0	0	3112	0.166	0.5774	0.983	0.094
New record	1	1	1	1	1	1	5223*				

Source: Author's calculations.

* Final occupational code assigned to the new record

However, considering information on the skills being demanded helps to identify a more precise sample of nearest neighbours in this example. More specifically, from Table G.5, it is possible to note that in the new record employers demand some skills, such as "Dispose waste," "Use food cutting tools," and "Clean surfaces." Moreover, in the training database, those skills are also demanded. The skills "Use food cutting tools" and "Clean surfaces" were mentioned in the first row, while the skill "Dispose waste" was mentioned in the third row. Consequently, it is possible to drop all those observations in the training database that do not have skills in common with the new record. By doing so, the last three rows in the training dataset are dropped (see Table G.6). As a result, Table G.6 presents a more precise training base for the algorithm. Indeed, the predicted code for the new record is 9412 "Kitchen assistant," which seems to be more accurate than 5223 "Shop sales assistants."

Table G.7 shows the accuracy level of each algorithm discussed in this book. The test database showed that the SVM algorithm in 60.4% of the cases (observations) correctly coded the job titles. The nearest neighbour algorithm that only used job titles had an accuracy level of 81.2%. However, the classification method with the highest accuracy level is the nearest neighbour algorithm that used both skills and job titles. In 92.3% of cases, this algorithm correctly coded the job titles. As shown above, this relatively high level of accuracy is because the nearest neighbour algorithm uses both the job title and information of skills, which helps to classify hard-coding observations.

Table G.6. **An extension of the nearest neighbour algorithm (Part 2)**

Source	Job title			Job description (skills)			Parameters				
	Technician	Assistant	Food	Dispose waste	Use food cutting tools	Clean surfaces	ISCO code	$p(c_i\|x)$	$s(x)$	$\left(\frac{K(X)}{K(X)+0.1}\right)$	$\Upsilon(c_i\|x)$
Training dataset	0	0	1	0	1	1	9412	0.75	0.5774	0.967	0.418
	0	0	1	1	0	1	9412				
	0	1	0	1	0	0	5223	0.25	0.5774	0.967	0.139
New record	1	1	1	1	1	1	9412*				

Source: Author's calculations.

* Final occupational code assigned to the new record.

Table G.7. **Comparison between the analysed classification methods**

Algorithm	Accuracy level
Supporting Vector Machine (SVM)	60.4%
Nearest neighbour algorithm using job titles	81.2%
Nearest neighbour algorithm using skills and job titles	92.3%

Source: Author's calculations based on vacancy and GEIH information, 2016-2018.

Appendix H: Additional Tables

Table H.1. **Occupations demanded in Colombia**

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
1	3322	Commercial sales representatives	878,503	15.4%
2	4223	Telephone switchboard operators	473,021	8.3%
3	4321	Stock clerks	472,076	8.3%
4	5223	Shop sales assistants	269,756	4.7%
5	5242	Sales demonstrators	235,481	4.1%
6	5230	Cashiers and ticket clerks	201,939	3.5%
7	4412	Mail carriers and sorting clerks	123,381	2.2%
8	5414	Security guards	111,717	2.0%
9	2411	Accountants	110,560	1.9%
10	1221	Sales and marketing managers	109,265	1.9%
11	4214	Debt-collectors and related workers	91,483	1.6%
12	9412	Kitchen helpers	75,535	1.3%
13	3343	Administrative and executive secretaries	73,364	1.3%
14	4110	General office clerks	69,875	1.2%
15	4322	Production clerks	67,997	1.2%
16	4311	Accounting and bookkeeping clerks	58,822	1.0%
17	8153	Sewing machine operators	54,628	1.0%
18	4222	Contact centre information clerks	50,337	0.9%
19	3312	Credit and loan officers	48,063	0.8%
20	5321	Health care assistants	45,279	0.8%
21	3115	Mechanical engineering technicians	40,808	0.7%
22	9333	Freight handlers	38,009	0.7%
23	4323	Transport clerks	37,532	0.7%
24	2635	Social work and counselling professionals	35,148	0.6%
25	3341	Office supervisors	34,921	0.6%
26	5131	Waiters	34,873	0.6%
27	3512	Information and communications technology user support technicians	33,977	0.6%
28	2221	Nursing professionals	33,337	0.6%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
29	8322	Car, taxi, and van drivers	31,782	0.6%
30	2412	Financial and investment advisers	31,424	0.5%
31	3511	Information and communications technology operations technicians	31,320	0.5%
32	1120	Managing directors and chief executives	31,212	0.5%
33	5221	Shopkeepers	30,940	0.5%
34	2634	Psychologists	30,926	0.5%
35	8343	Crane, hoist, and related plant operators	27,656	0.5%
36	2141	Industrial and production engineers	27,393	0.5%
37	7412	Electrical mechanics and fitters	26,181	0.5%
38	2211	Generalist medical practitioners	25,237	0.4%
39	7512	Bakers, pastry-cooks, and confectionery makers	24,275	0.4%
40	7119	Building frame and related trades workers not elsewhere classified	24,019	0.4%
41	4226	Receptionists (general)	22,728	0.4%
42	4415	Filing and copying clerks	22,465	0.4%
43	2166	Graphic and multimedia designers	21,913	0.4%
44	1324	Supply, distribution, and related managers	21,376	0.4%
45	3114	Electronics engineering technicians	21,346	0.4%
46	2619	Legal professionals not elsewhere classified	21,295	0.4%
47	3213	Pharmaceutical technicians and assistants	20,709	0.4%
48	9621	Messengers, package deliverers, and luggage porters	20,600	0.4%
49	2513	Web and multimedia developers	20,333	0.4%
50	3257	Environmental and occupational health inspectors and associates	19,814	0.3%
51	2151	Electrical engineers	21,238	0.4%
52	3435	Other artistic and cultural associate professionals	20,676	0.4%
53	2149	Engineering professionals not elsewhere classified	19,993	0.4%
54	9112	Cleaners and helpers in offices, hotels, and other establishments	18,385	0.3%
55	3323	Buyers	18,069	0.3%
56	5120	Cooks	17,489	0.3%
57	2511	Systems analysts	17,485	0.3%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
58	8332	Heavy truck and lorry drivers	17,401	0.3%
59	2431	Advertising and marketing professionals	17,179	0.3%
60	9622	Odd job persons	16,951	0.3%
61	7411	Building and related electricians	16,881	0.3%
62	9111	Domestic cleaners and helpers	16,751	0.3%
63	2523	Computer network professionals	16,659	0.3%
64	9329	Manufacturing labourers not elsewhere classified	15,209	0.3%
65	4227	Survey and market research interviewers	14,790	0.3%
66	7422	Information and communications technology installers and servicers	13,754	0.2%
67	2131	Biologists, botanists, zoologists, and related professionals	13,728	0.2%
68	7233	Agricultural and industrial machinery mechanics and repairers	13,724	0.2%
69	9313	Building construction labourers	13,449	0.2%
70	7322	Printers	13,409	0.2%
71	2161	Building architects	13,303	0.2%
72	3113	Electrical engineering technicians	13,285	0.2%
73	2142	Civil engineers	13,256	0.2%
74	2163	Product and garment designers	12,774	0.2%
75	2113	Chemists	12,428	0.2%
76	3112	Civil engineering technicians	12,180	0.2%
77	2212	Specialist medical practitioners	11,948	0.2%
78	8321	Motorcycle drivers	11,695	0.2%
79	2144	Mechanical engineers	11,678	0.2%
80	3122	Manufacturing supervisors	11,336	0.2%
81	2310	University and higher education teachers	11,291	0.2%
82	8142	Plastic products machine operators	10,969	0.2%
83	4313	Payroll clerks	10,835	0.2%
84	3334	Real estate agents and property managers	10,618	0.2%
85	7212	Welders and flame cutters	10,592	0.2%
86	2611	Lawyers	10,564	0.2%
87	2351	Education methods specialists	10,378	0.2%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
88	7115	Carpenters and joiners	10,311	0.2%
89	8160	Food and related products machine operators	10,204	0.2%
90	2152	Electronics engineers	10,175	0.2%
91	2261	Dentists	10,005	0.2%
92	3123	Construction supervisors	9,896	0.2%
93	3434	Chefs	9,781	0.2%
94	7231	Motor vehicle mechanics and repairers	9,514	0.2%
95	5142	Beauticians and related workers	8,901	0.2%
96	7221	Blacksmiths, hammersmiths, and forging press workers	8,861	0.2%
97	2262	Pharmacists	8,778	0.2%
98	2146	Mining engineers, metallurgists, and related professionals	8,507	0.2%
99	3339	Business services agents not elsewhere classified	8,491	0.2%
100	5222	Shop supervisors	8,277	0.1%
101	2622	Librarians and related information professionals	8,213	0.1%
102	2269	Health professionals not elsewhere classified	8,188	0.1%
103	2434	Information and communications technology sales professionals	8,116	0.1%
104	3522	Telecommunications engineering technicians	8,026	0.1%
105	9613	Sweepers and related labourers	7,888	0.1%
106	4416	Personnel clerks	7,795	0.1%
107	1439	Services managers not elsewhere classified	7,630	0.1%
108	6113	Gardeners, horticultural, and nursery growers	7,524	0.1%
109	8143	Paper products machine operators	7,420	0.1%
110	2433	Technical and medical sales professionals (excluding ICT)	7,394	0.1%
111	3111	Chemical and physical science technicians	7,223	0.1%
112	2330	Secondary education teachers	7,144	0.1%
113	7131	Painters and related workers	7,048	0.1%
114	2264	Physiotherapists	7,036	0.1%
115	9629	Elementary workers not elsewhere classified	6,792	0.1%
116	3321	Insurance representatives	6,724	0.1%
117	7421	Electronics mechanics and servicers	6,627	0.1%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
118	1212	Human resource managers	6,586	0.1%
119	3118	Draughtspersons	6,519	0.1%
120	7223	Metal working machine tool setters and operators	6,404	0.1%
121	1346	Financial and insurance services branch managers	6,269	0.1%
122	8183	Packing, bottling, and labelling machine operators	6,232	0.1%
123	5243	Door to door salespersons	6,228	0.1%
124	3411	Legal and related associate professionals	6,214	0.1%
125	7533	Sewing, embroidery, and related workers	6,174	0.1%
126	3521	Broadcasting and audio-visual technicians	5,974	0.1%
127	2265	Dieticians and nutritionists	5,734	0.1%
128	3117	Mining and metallurgical technicians	5,567	0.1%
129	7511	Butchers, fishmongers, and related food preparers	5,563	0.1%
130	5132	Bartenders	5,363	0.1%
131	5249	Sales workers not elsewhere classified	5,229	0.1%
132	4419	Clerical support workers not elsewhere classified	5,196	0.1%
133	2659	Creative and performing artists not elsewhere classified	5,068	0.1%
134	7323	Print finishing and binding workers	5,041	0.1%
135	2120	Mathematicians, actuaries, and statisticians	4,981	0.1%
136	8131	Chemical products plant and machine operators	4,960	0.1%
137	1323	Construction managers	4,918	0.1%
138	1321	Manufacturing managers	4,895	0.1%
139	5322	Home-based personal care workers	4,837	0.1%
140	2512	Software developers	4,740	0.1%
141	2353	Other language teachers	4,710	0.1%
142	4212	Bookmakers, croupiers, and related gaming workers	4,577	0.1%
143	7127	Air conditioning and refrigeration mechanics	4,537	0.1%
144	5329	Personal care workers in health services not elsewhere classified	4,451	0.1%
145	2153	Telecommunications engineers	4,385	0.1%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
146	8342	Earthmoving and related plant operators	4,365	0.1%
147	5211	Stall and market salespersons	4,320	0.1%
148	1420	Retail and wholesale trade managers	4,300	0.1%
149	5244	Contact centre salespersons	4,041	0.1%
150	3422	Sports coaches, instructors, and officials	3,806	0.1%
151	2250	Veterinarians	3,647	0.1%
152	7318	Handicraft workers in textile, leather, and related materials	3,627	0.1%
153	2521	Database designers and administrators	3,588	0.1%
154	7321	Pre-press technicians	3,558	0.1%
155	3412	Social work associate professionals	3,488	0.1%
156	8156	Shoemaking and related machine operators	3,463	0.1%
157	7311	Precision-instrument makers and repairers	3,371	0.1%
158	5311	Child care workers	3,364	0.1%
159	5153	Building caretakers	3,360	0.1%
160	3211	Medical imaging and therapeutic equipment technicians	3,337	0.1%
161	5164	Pet groomers and animal care workers	3,174	0.1%
162	3333	Employment agents and contractors	3,169	0.1%
163	2145	Chemical engineers	3,144	0.1%
164	2143	Environmental engineers	3,104	0.1%
165	3514	Web technicians	3,104	0.1%
166	2132	Farming, forestry, and fisheries advisers	2,957	0.1%
167	7111	House builders	2,937	0.1%
168	2164	Town and traffic planners	2,755	0.05%
169	5112	Transport conductors	2,728	0.05%
170	1330	Information and communications technology service managers	2,652	0.05%
171	2359	Teaching professionals not elsewhere classified	2,611	0.05%
172	8341	Mobile farm and forestry plant operators	2,587	0.05%
173	7536	Shoemakers and related workers	2,587	0.05%
174	7132	Spray painters and varnishers	2,571	0.05%
175	4221	Travel consultants and clerks	2,563	0.05%
176	3313	Accounting associate professionals	2,488	0.04%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
177	9520	Street vendors (excluding food)	2,471	0.04%
178	1411	Hotel managers	2,414	0.04%
179	3212	Medical and pathology laboratory technicians	2,225	0.04%
180	3142	Agricultural technicians	2,186	0.04%
181	2352	Special needs teachers	2,127	0.04%
182	7222	Toolmakers and related workers	2,042	0.04%
183	8181	Glass and ceramics plant operators	1,995	0.04%
184	5163	Undertakers and embalmers	1,973	0.04%
185	2643	Translators, interpreters, and other linguists	1,946	0.03%
186	7514	Fruit, vegetable, and related preservers	1,913	0.03%
187	7531	Tailors, dressmakers, furriers, and hatters	1,875	0.03%
188	5152	Domestic housekeepers	1,855	0.03%
189	7534	Upholsterers and related workers	1,836	0.03%
190	5111	Travel attendants and travel stewards	1,832	0.03%
191	4229	Client information workers not elsewhere classified	1,812	0.03%
192	7211	Metal moulders and coremakers	1,799	0.03%
193	3513	Computer network and systems technicians	1,696	0.03%
194	7114	Concrete placers, concrete finishers, and related workers	1,667	0.03%
195	3153	Aircraft pilots and related associate professionals	1,657	0.03%
196	2413	Financial analysts	1,647	0.03%
197	2633	Philosophers, historians, and political scientists	1,623	0.03%
198	2133	Environmental protection professionals	1,615	0.03%
199	7112	Bricklayers and related workers	1,597	0.03%
200	5419	Protective services workers not elsewhere classified	1,574	0.03%
201	3352	Government tax and excise officials	1,573	0.03%
202	8159	Textile, fur, and leather products machine operators not elsewhere classified	1,567	0.03%
203	2267	Optometrists and ophthalmic opticians	1,557	0.03%
204	7213	Sheet-metal workers	1,551	0.03%
205	9211	Crop farm labourers	1,514	0.03%
206	5141	Hairdressers	1,507	0.03%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
207	9411	Fast food preparers	1,485	0.03%
208	8219	Assemblers not elsewhere classified	1,484	0.03%
209	8121	Metal processing plant operators	1,455	0.03%
210	2165	Cartographers and surveyors	1,434	0.03%
211	1211	Finance managers	1,421	0.03%
212	3354	Government licensing officials	1,408	0.03%
213	3133	Chemical processing plant controllers	1,405	0.02%
214	7315	Glass makers, cutters, grinders, and finishers	1,399	0.02%
215	2641	Authors and related writers	1,398	0.02%
216	2421	Management and organization analysts	1,395	0.02%
217	7532	Garment and related pattern-makers and cutters	1,380	0.02%
218	2631	Economists	1,317	0.02%
219	3432	Interior designers and decorators	1,312	0.02%
220	7544	Fumigators and other pest and weed controllers	1,245	0.02%
221	3154	Air traffic controllers	1,234	0.02%
222	2642	Journalists	1,234	0.02%
223	3152	Ships' deck officers and pilots	1,223	0.02%
224	3431	Photographers	1,200	0.02%
225	2423	Personnel and careers professionals	1,198	0.02%
226	8157	Laundry machine operators	1,157	0.02%
227	7314	Potters and related workers	1,112	0.02%
228	1349	Professional services managers not elsewhere classified	1,103	0.02%
229	7522	Cabinet-makers and related workers	1,099	0.02%
230	9612	Refuse sorters	1,065	0.02%
231	2632	Sociologists, anthropologists, and related professionals	1,062	0.02%
232	4132	Data entry clerks	1,036	0.02%
233	2656	Announcers on radio, television, and other media	1,016	0.02%
234	3314	Statistical, mathematical, and related associate professionals	1,011	0.02%
235	5245	Service station attendants	1,002	0.02%
236	3331	Clearing and forwarding agents	998	0.02%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
237	7122	Floor layers and tile setters	969	0.02%
238	2432	Public relations professionals	966	0.02%
239	3132	Incinerator and water treatment plant operators	948	0.02%
240	1114	Senior officials of special-interest organizations	945	0.02%
241	2114	Geologists and geophysicists	944	0.02%
242	2266	Audiologists and speech therapists	939	0.02%
243	3423	Fitness and recreation instructors and program leaders	939	0.02%
244	8152	Weaving and knitting machine operators	936	0.02%
245	8114	Cement, stone, and other mineral products machine operators	925	0.02%
246	8331	Bus and tram drivers	901	0.02%
247	7125	Glaziers	895	0.02%
248	1345	Education managers	876	0.02%
249	3332	Conference and event planners	864	0.02%
250	7224	Metal polishers, wheel grinders, and tool sharpeners	829	0.01%
251	2342	Early childhood educators	820	0.01%
252	9623	Meter readers and vending-machine collectors	811	0.01%
253	2529	Database and network professionals not elsewhere classified	803	0.01%
254	7513	Dairy-products makers	797	0.01%
255	3255	Physiotherapy technicians and assistants	791	0.01%
256	1412	Restaurant managers	787	0.01%
257	9510	Street and related service workers	783	0.01%
258	2654	Film, stage, and related directors and producers	760	0.01%
259	3155	Air traffic safety electronics technicians	741	0.01%
260	3214	Medical and dental prosthetic technicians	736	0.01%
261	5411	Fire-fighters	733	0.01%
262	7214	Structural-metal preparers and erectors	731	0.01%
263	7313	Jewellery and precious-metal workers	727	0.01%
264	8312	Railway brake, signal, and switch operators	720	0.01%
265	1341	Child care services managers	700	0.01%
266	7413	Electrical line installers and repairers	684	0.01%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
267	2519	Software and applications developers and analysts not elsewhere classified	683	0.01%
268	8182	Steam engine and boiler operators	675	0.01%
269	8344	Lifting truck operators	671	0.01%
270	3251	Dental assistants and therapists	660	0.01%
271	4213	Pawnbrokers and money-lenders	652	0.01%
272	3254	Dispensing opticians	640	0.01%
273	4225	Enquiry clerks	630	0.01%
274	8151	Fibre preparing, spinning, and winding machine operators	583	0.01%
275	5169	Personal services workers not elsewhere classified	558	0.01%
276	2424	Training and staff development professionals	547	0.01%
277	2653	Dancers and choreographers	546	0.01%
278	3342	Legal secretaries	523	0.01%
279	9121	Hand launderers and pressers	504	0.01%
280	7126	Plumbers and pipe fitters	500	0.01%
281	7124	Insulation workers	495	0.01%
282	1222	Advertising and public relations managers	493	0.01%
283	5165	Driving instructors	474	0.01%
284	3139	Process control technicians not elsewhere classified	468	0.01%
285	2621	Archivists and curators	465	0.01%
286	1431	Sports, recreation, and cultural centre managers	463	0.01%
287	8350	Ships' deck crews and related workers	461	0.01%
288	9214	Garden and horticultural labourers	442	0.01%
289	3259	Health associate professionals not elsewhere classified	441	0.01%
290	8211	Mechanical machinery assemblers	434	0.01%
291	8154	Bleaching, dyeing, and fabric cleaning machine operators	426	0.01%
292	4411	Library clerks	393	0.01%
293	7215	Riggers and cable splicers	381	0.01%
294	5246	Food service counter attendants	364	0.01%
295	3143	Forestry technicians	362	0.01%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
296	8155	Fur and leather preparing machine operators	361	0.01%
297	2522	Systems administrators	360	0.01%
298	7317	Handicraft workers in wood, basketry, and related materials	352	0.01%
299	1219	Business services and administration managers not elsewhere classified	352	0.01%
300	3151	Ships' engineers	344	0.01%
301	7316	Sign writers, decorative painters, engravers, and etchers	337	0.01%
302	7516	Tobacco preparers and tobacco products makers	330	0.01%
303	6121	Livestock and dairy producers	325	0.01%
304	5312	Teachers' aides	308	0.01%
305	2612	Judges	297	0.01%
306	1322	Mining managers	285	0.01%
307	2320	Vocational education teachers	282	0.01%
308	2636	Religious professionals	258	0.005%
309	3240	Veterinary technicians and assistants	257	0.005%
310	4312	Statistical, finance, and insurance clerks	256	0.005%
311	7121	Roofers	251	0.004%
312	4211	Bank tellers and related clerks	249	0.004%
313	7541	Underwater divers	242	0.004%
314	3121	Mining supervisors	241	0.004%
315	8111	Miners and quarriers	237	0.004%
316	3141	Life science technicians (excluding medical)	236	0.004%
317	4131	Typists and word processing operators	218	0.004%
318	7535	Pelt dressers, tanners, and fellmongers	216	0.004%
319	7232	Aircraft engine mechanics and repairers	208	0.004%
320	3353	Government social benefits officials	206	0.004%
321	6112	Tree and shrub crop growers	196	0.004%
322	8171	Pulp and papermaking plant operators	196	0.004%
323	3134	Petroleum and natural gas refining plant operators	189	0.003%
324	2356	Information technology trainers	188	0.003%
325	3355	Police inspectors and detectives	180	0.003%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
326	2222	Midwifery professionals	180	0.003%
327	8141	Rubber products machine operators	179	0.003%
328	2514	Applications programmers	171	0.003%
329	7113	Stonemasons, stone cutters, splitters, and carvers	168	0.003%
330	8122	Metal finishing, plating, and coating machine operators	167	0.003%
331	8132	Photographic products machine operators	159	0.003%
332	9611	Garbage and recycling collectors	125	0.002%
333	6210	Forestry and related workers	124	0.002%
334	5151	Cleaning and housekeeping supervisors in offices, hotels, and other establishments	122	0.002%
335	8112	Mineral and stone processing plant operators	109	0.002%
336	4413	Coding, proof-reading, and related clerks	109	0.002%
337	8172	Wood processing plant operators	98	0.002%
338	2655	Actors	94	0.002%
339	7234	Bicycle and related repairers	93	0.002%
340	3258	Ambulance workers	88	0.002%
341	9321	Hand packers	86	0.002%
342	4224	Hotel receptionists	81	0.001%
343	3135	Metal production process controllers	75	0.001%
344	8311	Locomotive engine drivers	75	0.001%
345	3324	Trade brokers	71	0.001%
346	9334	Shelf fillers	67	0.001%
347	9212	Livestock farm labourers	65	0.001%
348	4120	Secretaries (general)	63	0.001%
349	1343	Aged care services managers	60	0.001%
350	3359	Regulatory government associate professionals not elsewhere classified	60	0.001%
351	2341	Primary school teachers	59	0.001%
352	1311	Agricultural and forestry production managers	58	0.001%
353	2355	Other arts teachers	56	0.001%
354	9312	Civil engineering labourers	56	0.001%
355	5113	Travel guides	48	0.001%
356	1342	Health services managers	47	0.001%

Position	ISCO-08 code	Occupation	Number of jobs	Percentage
357	1223	Research and development managers	43	0.001%
358	6221	Aquaculture workers	41	0.001%
359	6111	Field crop and vegetable growers	34	0.001%
360	3344	Medical secretaries	34	0.001%
361	2354	Other music teachers	31	0.001%
362	9216	Fishery and aquaculture labourers	26	0.0005%
363	7521	Wood treaters	21	0.0004%
364	2651	Visual artists	17	0.0003%
365	6123	Apiarists and sericulturists	14	0.0003%
366	2112	Meteorologists	13	0.0002%
367	8113	Well drillers and borers and related workers	11	0.0002%
368	8189	Stationary plant and machine operators not elsewhere classified	10	0.0002%
369	9129	Other cleaning workers	10	0.0002%
370	7319	Handicraft workers not elsewhere classified	10	0.0002%
371	5412	Police officers	8	0.0001%
372	2162	Landscape architects	8	0.0001%
373	1213	Policy and planning managers	7	0.0001%
374	9311	Mining and quarrying labourers	6	0.0001%
375	5212	Street food salespersons	5	0.0001%

Source: Author's calculations based on vacancy information, 2016-2018.

Table H.2. **Occupational distribution of Colombian workers**

#	ISCO title	Formal workers	ISCO title	Informal workers
1	Sales demonstrators	4.8%	Sales demonstrators	16.4%
2	Secondary education teachers	4.5%	Domestic cleaners and helpers	6.0%
3	Security guards	3.7%	Car, taxi, and van drivers	6.0%
4	Cleaners and helpers in offices, hotels, and other establishments	3.6%	Stall and market salespersons	3.7%
5	Car, taxi, and van drivers	3.0%	Cleaners and helpers in offices, hotels, and other establishments	3.3%
6	Stock clerks	2.0%	Cooks	2.9%

#	ISCO title	Formal workers	ISCO title	Informal workers
7	Health care assistants	1.9%	Commercial sales representatives	2.3%
8	Building and related electricians	1.8%	Bricklayers and related workers	2.1%
9	Accounting and bookkeeping clerks	1.7%	Child care workers	2.1%
10	Waiters	1.5%	Building and related electricians	1.9%
11	Welders and flame cutters	1.5%	Beauticians and related workers	1.9%
12	Primary school teachers	1.5%	Sewing machine operators	1.9%
13	Child care workers	1.5%	Services managers not elsewhere classified	1.8%
14	Sewing machine operators	1.4%	Shop keepers	1.8%
15	Mail carriers and sorting clerks	1.3%	Crop farm labourers	1.7%
16	Cooks	1.3%	Motorcycle drivers	1.6%
17	Cashiers and ticket clerks	1.3%	Motor vehicle mechanics and repairers	1.6%
18	Contact centre information clerks	1.1%	Construction supervisors	1.4%
19	Kitchen helpers	1.0%	Freight handlers	1.2%
20	Senior officials of special-interest organizations	1.0%	Waiters	1.2%
21	Lawyers	1.0%	Kitchen helpers	1.2%
22	Services managers not elsewhere classified	1.0%	Tailors, dressmakers, furriers, and hatters	1.1%
23	Heavy truck and lorry drivers	1.0%	Painters and related workers	1.1%
24	Commercial sales representatives	1.0%	Heavy truck and lorry drivers	1.0%
25	Policy administration professionals	1.0%	Bakers, pastry-cooks, and confectionery makers	0.9%
26	Police inspectors and detectives	0.9%	House builders	0.9%
27	Generalist medical practitioners	0.8%	Door to door salespersons	0.9%
28	Agricultural and industrial machinery mechanics and repairers	0.8%	Real estate agents and property managers	0.8%
29	Administrative and executive secretaries	0.8%	Security guards	0.7%
30	Office supervisors	0.8%	Shop sales assistants	0.7%
31	Human resource managers	0.7%	Cabinet-makers and related workers	0.7%

#	ISCO title	Formal workers	ISCO title	Informal workers
32	Financial and insurance services branch managers	0.7%	Fast food preparers	0.7%
33	Transport clerks	0.7%	Butchers, fishmongers, and related food preparers	0.7%
34	Nursing professionals	0.7%	Hairdressers	0.7%
35	Supply, distribution, and related managers	0.7%	Messengers, package deliverers, and luggage porters	0.6%
36	House builders	0.6%	Shoemakers and related workers	0.6%
37	Physical and engineering science technicians not elsewhere classified	0.6%	Handicraft workers in textile, leather, and related materials	0.6%
38	Retail and wholesale trade managers	0.6%	Welders and flame cutters	0.6%
39	Civil engineers	0.6%	Carpenters and joiners	0.6%
40	Freight handlers	0.6%	Gardeners, horticultural, and nursery growers	0.6%
41	Shop sales assistants	0.6%	Mail carriers and sorting clerks	0.6%
42	Secretaries (general)	0.6%	Livestock and dairy producers	0.6%
43	Construction supervisors	0.6%	Food service counter attendants	0.5%
44	Graphic and multimedia designers	0.6%	Food and related products machine operators	0.5%
45	Butchers, fishmongers, and related food preparers	0.6%	Retail and wholesale trade managers	0.5%
46	Bakers, pastry-cooks, and confectionery makers	0.6%	Field crop and vegetable growers	0.4%
47	Systems analysts	0.6%	Stonemasons, stone cutters, splitters, and carvers	0.4%
48	Motor vehicle mechanics and repairers	0.5%	Contact centre salespersons	0.4%
49	Crop farm labourers	0.5%	Mobile farm and forestry plant operators	0.4%
50	Bricklayers and related workers	0.5%	Miners and quarriers	0.4%
51	Crane, hoist, and related plant operators	0.5%	Sewing, embroidery, and related workers	0.4%
52	Messengers, package deliverers, and luggage porters	0.5%	Hand launderers and pressers	0.4%
53	Information and communications technology installers and servicers	0.5%	Fruit, vegetable, and related preservers	0.4%

#	ISCO title	Formal workers	ISCO title	Informal workers
54	Legal professionals not elsewhere classified	0.5%	Cashiers and ticket clerks	0.3%
55	Industrial and production engineers	0.5%	Database and network professionals not elsewhere classified	0.3%
56	Driving instructors	0.5%	Restaurant managers	0.3%
57	Psychologists	0.5%	Electrical mechanics and fitters	0.3%
58	Specialist medical practitioners	0.4%	Tree and shrub crop growers	0.3%
59	Real estate agents and property managers	0.4%	Bookmakers, croupiers, and related gaming workers	0.3%
60	Musicians, singers, and composers	0.4%	Building caretakers	0.3%
61	Teaching professionals not elsewhere classified	0.4%	Home-based personal care workers	0.3%
62	Business services agents not elsewhere classified	0.4%	Bartenders	0.3%
63	Accountants	0.4%	Stock clerks	0.2%
64	Electrical mechanics and fitters	0.4%	Senior officials of special-interest organizations	0.2%
65	Filing and copying clerks	0.4%	Packing, bottling, and labelling machine operators	0.2%
66	Painters and related workers	0.4%	Sheet-metal workers	0.2%
67	University and higher education teachers	0.4%	Agricultural and industrial machinery mechanics and repairers	0.2%
68	Door to door salespersons	0.4%	Laundry machine operators	0.2%
69	Mathematicians, actuaries, and statisticians	0.4%	Garbage and recycling collectors	0.2%
70	Securities and finance dealers and brokers	0.4%	Transport clerks	0.2%
71	Plastic products machine operators	0.4%	Other artistic and cultural associate professionals	0.2%
72	Other language teachers	0.3%	Manufacturing supervisors	0.2%
73	Packing, bottling, and labelling machine operators	0.3%	Fumigators and other pest and weed controllers	0.2%
74	Insurance representatives	0.3%	Spray painters and varnishers	0.2%
75	Paper products machine operators	0.3%	Potters and related workers	0.2%
76	Building architects	0.3%	Street vendors (excluding food)	0.2%

#	ISCO title	Formal workers	ISCO title	Informal workers
77	Civil engineering technicians	0.3%	Accounting and bookkeeping clerks	0.2%
78	Dentists	0.3%	Buyers	0.2%
79	Receptionists (general)	0.3%	Jewellery and precious-metal workers	0.2%
80	Religious professionals	0.3%	Product and garment designers	0.2%
81	Sales and marketing managers	0.3%	Shoemaking and related machine operators	0.2%
82	Buyers	0.3%	Street food salespersons	0.2%
83	Manufacturing managers	0.3%	Assemblers not elsewhere classified	0.1%
84	Food and related products machine operators	0.3%	Refuse sorters	0.1%
85	Social work and counselling professionals	0.3%	Upholsterers and related workers	0.1%
86	Restaurant managers	0.3%	Civil engineers	0.1%
87	Construction managers	0.3%	Credit and loans officers	0.1%
88	Environmental and occupational health inspectors and associates	0.3%	Teachers' aides	0.1%
89	Electrical engineers	0.3%	Toolmakers and related workers	0.1%
90	Carpenters and joiners	0.3%	Translators, interpreters, and other linguists	0.1%
91	Information and communications technology sales professionals	0.3%	Administrative and executive secretaries	0.1%
92	Translators, interpreters, and other linguists	0.3%	Garment and related pattern-makers and cutters	0.1%
93	Bus and tram drivers	0.3%	Policy administration professionals	0.1%
94	Pharmaceutical technicians and assistants	0.3%	Glass and ceramics plant operators	0.1%
95	Beauticians and related workers	0.3%	Information and communications technology installers and servicers	0.1%
96	Police officers	0.3%	Aquaculture workers	0.1%
97	Cabinet-makers and related workers	0.2%	Civil engineering technicians	0.1%
98	Gardeners, horticultural, and nursery growers	0.2%	Physical and engineering science technicians not elsewhere classified	0.1%

#	ISCO title	Formal workers	ISCO title	Informal workers
99	Electrical engineering technicians	0.2%	Bus and tram drivers	0.1%
100	Advertising and marketing professionals	0.2%	Crane, hoist, and related plant operators	0.1%
101	Statistical, finance, and insurance clerks	0.2%	Electronics mechanics and servicers	0.1%
102	Meter readers and vending-machine collectors	0.2%	Building construction labourers	0.1%
103	Stall and market salespersons	0.2%	Secretaries (general)	0.1%
104	Glass and ceramics plant operators	0.2%	Poultry producers	0.1%
105	Manufacturing supervisors	0.2%	Pawnbrokers and money-lenders	0.1%
106	Enquiry clerks	0.2%	Pet groomers and animal care workers	0.1%
107	Financial analysts	0.2%	Primary school teachers	0.1%
108	Information and communications technology user support technicians	0.2%	Secondary education teachers	0.1%
109	Other artistic and cultural associate professionals	0.2%	Dairy-products makers	0.1%
110	Personal care workers in health services not elsewhere classified	0.2%	Vehicle cleaners	0.1%
111	Broadcasting and audio-visual technicians	0.2%	Sales workers not elsewhere classified	0.1%
112	Food service counter attendants	0.2%	Personal care workers in health services not elsewhere classified	0.1%
113	Biologists, botanists, zoologists, and related professionals	0.2%	Health care assistants	0.1%
114	Telecommunications engineers	0.2%	Debt-collectors and related workers	0.1%
115	Toolmakers and related workers	0.2%	Handicraft workers in wood, basketry, and related materials	0.1%
116	Chefs	0.2%	Bicycle and related repairers	0.1%
117	Early childhood educators	0.2%	Clearing and forwarding agents	0.1%
118	Information and communications technology service managers	0.2%	Companions and valets	0.1%
119	Bank tellers and related clerks	0.2%	Biologists, botanists, zoologists, and related professionals	0.1%

#	ISCO title	Formal workers	ISCO title	Informal workers
120	Fitness and recreation instructors and program leaders	0.2%	Air conditioning and refrigeration mechanics	0.1%
121	Assemblers not elsewhere classified	0.2%	Hotel receptionists	0.1%
122	Management and organization analysts	0.2%	Graphic and multimedia designers	0.1%
123	Photographers	0.2%	Dentists	0.1%
124	Armed forces occupations, other ranks	0.2%	Insurance representatives	0.1%
125	Electronics mechanics and servicers	0.2%	Livestock farm labourers	0.1%
126	Stonemasons, stone cutters, splitters, and carvers	0.2%	Industrial and production engineers	0.1%
127	Physiotherapists	0.2%	Concrete placers, concrete finishers, and related workers	0.1%
128	Shoemakers and related workers	0.2%	Business services agents not elsewhere classified	0.1%
129	Medical and pathology laboratory technicians	0.2%	Fishery and aquaculture labourers	0.1%
130	Archivists and curators	0.2%	Receptionists (general)	0.1%
131	Mechanical engineering technicians	0.2%	Fitness and recreation instructors and program leaders	0.1%
132	Fruit, vegetable, and related preservers	0.2%	Broadcasting and audio-visual technicians	0.1%
133	Agricultural technicians	0.2%	Floor layers and tile setters	0.1%
134	Air conditioning and refrigeration mechanics	0.2%	Print finishing and binding workers	0.1%
135	Chemical and physical science technicians	0.2%	Chefs	0.1%
136	Shoemaking and related machine operators	0.2%	Medical and dental prosthetic technicians	0.1%
137	Engineering professionals not elsewhere classified	0.2%	Religious professionals	0.1%
138	Product and garment designers	0.2%	Sports coaches, instructors, and officials	0.1%
139	Sports coaches, instructors, and officials	0.2%	Paper products machine operators	0.1%
140	Sweepers and related labourers	0.2%	Supply, distribution, and related managers	0.1%
141	Electronics engineers	0.2%	Printers	0.1%

#	ISCO title	Formal workers	ISCO title	Informal workers
142	Financial and investment advisers	0.2%	Lawyers	0.1%
143	Bartenders	0.2%	Plastic products machine operators	0.1%
144	Information and communications technology operations technicians	0.2%	Managing directors and chief executives	0.1%
145	Education methods specialists	0.2%	Bleaching, dyeing, and fabric cleaning machine operators	0.1%
146	Electronics engineering technicians	0.2%	Drivers of animal-drawn vehicles and machinery	0.1%
147	Managing directors and chief executives	0.1%	Enquiry clerks	0.1%
148	Survey and market research interviewers	0.1%	Conference and event planners	0.1%
149	Legislators	0.1%	Advertising and marketing professionals	0.1%
150	Mechanical engineers	0.1%	Legal professionals not elsewhere classified	0.1%
151	Debt-collectors and related workers	0.1%	Lifting truck operators	0.1%
152	Building construction labourers	0.1%	Medical and pathology laboratory technicians	0.0%
153	Metal working machine tool setters and operators	0.1%	Electrical engineering technicians	0.0%
154	Journalists	0.1%	Metal working machine tool setters and operators	0.0%
155	Hand packers	0.1%	Electrical engineers	0.0%
156	Bookmakers, croupiers, and related gaming workers	0.1%	Elementary workers not elsewhere classified	0.0%
157	Building caretakers	0.1%	Musicians, singers, and composers	0.0%
158	Payroll clerks	0.1%	Metal processing plant operators	0.0%
159	Personnel and careers professionals	0.1%	Manufacturing managers	0.0%
160	Database and network professionals not elsewhere classified	0.1%	Precision-instrument makers and repairers	0.0%
161	Plumbers and pipe fitters	0.1%	Pre-press technicians	0.0%
162	Fast food preparers	0.1%	Police inspectors and detectives	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
163	Contact centre salespersons	0.1%	Hotel managers	0.0%
164	Miners and quarriers	0.1%	Mechanical engineering technicians	0.0%
165	Chemists	0.1%	Forestry and related workers	0.0%
166	Special needs teachers	0.1%	Photographers	0.0%
167	Earthmoving and related plant operators	0.1%	Agricultural and forestry production managers	0.0%
168	Fumigators and other pest and weed controllers	0.1%	Armed forces occupations, other ranks	0.0%
169	Technical and medical sales professionals (excluding ICT)	0.1%	Meteorologists	0.0%
170	Hand launderers and pressers	0.1%	Sales and marketing managers	0.0%
171	Veterinarians	0.1%	Generalist medical practitioners	0.0%
172	Hotel receptionists	0.1%	Driving instructors	0.0%
173	Accounting associate professionals	0.1%	Personal services workers not elsewhere classified	0.0%
174	Health services managers	0.1%	Tobacco preparers and tobacco products makers	0.0%
175	Teachers' aides	0.1%	Veterinarians	0.0%
176	Personal services workers not elsewhere classified	0.1%	Construction managers	0.0%
177	Finance managers	0.1%	Technical and medical sales professionals (excluding ICT)	0.0%
178	Hairdressers	0.1%	Weaving and knitting machine operators	0.0%
179	Bleaching, dyeing, and fabric cleaning machine operators	0.1%	Deep-sea fishery workers	0.0%
180	Telephone switchboard operators	0.1%	Financial and insurance services branch managers	0.0%
181	Librarians and related information professionals	0.1%	Mineral and stone processing plant operators	0.0%
182	Spray painters and varnishers	0.1%	Inland and coastal waters fishery workers	0.0%
183	Fire-fighters	0.1%	Structural-metal preparers and erectors	0.0%
184	Elementary workers not elsewhere classified	0.1%	Cement, stone, and other mineral products machine operators	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
185	Incinerator and water treatment plant operators	0.1%	Plumbers and pipe fitters	0.0%
186	Field crop and vegetable growers	0.1%	Chemical products plant and machine operators	0.0%
187	Garbage and recycling collectors	0.1%	Pharmaceutical technicians and assistants	0.0%
188	Telecommunications engineering technicians	0.1%	Sign writers, decorative painters, engravers, and etchers	0.0%
189	Metal processing plant operators	0.1%	Telephone switchboard operators	0.0%
190	Home-based personal care workers	0.1%	Contact centre information clerks	0.0%
191	Database designers and administrators	0.1%	Specialist medical practitioners	0.0%
192	Cartographers and surveyors	0.1%	Incinerator and water treatment plant operators	0.0%
193	Handicraft workers in textile, leather, and related materials	0.1%	Wood processing plant operators	0.0%
194	Chemical engineering technicians	0.1%	Blacksmiths, hammersmiths, and forging press workers	0.0%
195	Service station attendants	0.1%	Securities and finance dealers and brokers	0.0%
196	Travel consultants and clerks	0.1%	Steam engine and boiler operators	0.0%
197	Mining and metallurgical technicians	0.1%	Transport conductors	0.0%
198	Medical imaging and therapeutic equipment technicians	0.1%	Teaching professionals not elsewhere classified	0.0%
199	Social work associate professionals	0.1%	Other arts teachers	0.0%
200	Mining supervisors	0.1%	Metal polishers, wheel grinders, and tool sharpeners	0.0%
201	Structural-metal preparers and erectors	0.1%	Building architects	0.0%
202	Farming, forestry, and fisheries advisers	0.1%	Aged care services managers	0.0%
203	Building frame and related trades workers not elsewhere classified	0.1%	Physiotherapy technicians and assistants	0.0%
204	Conference and event planners	0.1%	Fur and leather preparing machine operators	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
205	Sheet-metal workers	0.1%	Metal finishing, plating, and coating machine operators	0.0%
206	Upholsterers and related workers	0.1%	Sports, recreation, and cultural centre managers	0.0%
207	Credit and loans officers	0.1%	Early childhood educators	0.0%
208	Precision-instrument makers and repairers	0.1%	Fire-fighters	0.0%
209	Judges	0.1%	Hand packers	0.0%
210	Print finishing and binding workers	0.1%	Travel guides	0.0%
211	Livestock and dairy producers	0.1%	Office supervisors	0.0%
212	Environmental engineers	0.1%	Service station attendants	0.0%
213	Garment and related pattern-makers and cutters	0.1%	Chemical and physical science technicians	0.0%
214	Mining engineers, metallurgists, and related professionals	0.1%	Building frame and related trades workers not elsewhere classified	0.0%
215	Authors and related writers	0.1%	Meter readers and vending-machine collectors	0.0%
216	Shop keepers	0.1%	Accountants	0.0%
217	Livestock farm labourers	0.1%	Agricultural technicians	0.0%
218	Draughtspersons	0.1%	Domestic housekeepers	0.0%
219	Non-commissioned armed forces officers	0.1%	Information and communications technology sales professionals	0.0%
220	Electrical line installers and repairers	0.1%	Nursing professionals	0.0%
221	Professional services managers not elsewhere classified	0.1%	Chemical engineers	0.0%
222	Pharmacists	0.1%	Chemists	0.0%
223	Jewellery and precious-metal workers	0.1%	Manufacturing labourers not elsewhere classified	0.0%
224	Aircraft pilots and related associate professionals	0.1%	Human resource managers	0.0%
225	Mineral and stone processing plant operators	0.1%	Electrical line installers and repairers	0.0%
226	Dental assistants and therapists	0.1%	Glaziers	0.0%
227	Client information workers not elsewhere classified	0.1%	Pharmacists	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
228	Medical and dental prosthetic technicians	0.1%	Roofers	0.0%
229	Cleaning and housekeeping supervisors in offices, hotels, and other establishments	0.1%	Social work and counselling professionals	0.0%
230	Vehicle cleaners	0.1%	Interior designers and decorators	0.0%
231	Civil engineering labourers	0.1%	Environmental and occupational health inspectors and associates	0.0%
232	Business services and administration managers not elsewhere classified	0.1%	Financial and investment advisers	0.0%
233	Chemical products plant and machine operators	0.1%	Fibre preparing, spinning, and winding machine operators	0.0%
234	Economists	0.1%	Electronics engineering technicians	0.0%
235	Tailors, dressmakers, furriers, and hatters	0.1%	Education methods specialists	0.0%
236	Education managers	0.0%	Textile, fur, and leather products machine operators not elsewhere classified	0.0%
237	Pre-press technicians	0.0%	Civil engineering labourers	0.0%
238	Interior designers and decorators	0.0%	Street and related service workers	0.0%
239	Printers	0.0%	Social work associate professionals	0.0%
240	Health professionals not elsewhere classified	0.0%	Engineering professionals not elsewhere classified	0.0%
241	Chemical engineers	0.0%	Systems analysts	0.0%
242	Employment agents and contractors	0.0%	Travel consultants and clerks	0.0%
243	Commissioned armed forces officers	0.0%	Insulation workers	0.0%
244	Riggers and cable splicers	0.0%	Dispensing opticians	0.0%
245	Wood processing plant operators	0.0%	Mechanical engineers	0.0%
246	Poultry producers	0.0%	Forestry labourers	0.0%
247	Metal finishing, plating, and coating machine operators	0.0%	Archivists and curators	0.0%
248	Aquaculture and fisheries production managers	0.0%	Glass makers, cutters, grinders, and finishers	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
249	Weaving and knitting machine operators	0.0%	Cleaning and housekeeping supervisors in offices, hotels, and other establishments	0.0%
250	Floor layers and tile setters	0.0%	Plasterers	0.0%
251	Companions and valets	0.0%	Chemical engineering technicians	0.0%
252	Laundry machine operators	0.0%	Creative and performing artists not elsewhere classified	0.0%
253	Railway brake, signal, and switch operators	0.0%	Aircraft pilots and related associate professionals	0.0%
254	Street vendors (excluding food)	0.0%	Cartographers and surveyors	0.0%
255	Environmental protection professionals	0.0%	Advertising and public relations managers	0.0%
256	Transport conductors	0.0%	Pelt dressers, tanners, and fellmongers	0.0%
257	Announcers on radio, television, and other media	0.0%	Payroll clerks	0.0%
258	Mobile farm and forestry plant operators	0.0%	Dental assistants and therapists	0.0%
259	Process control technicians not elsewhere classified	0.0%	Web and multimedia developers	0.0%
260	Aircraft engine mechanics and repairers	0.0%	Draughtspersons	0.0%
261	Library clerks	0.0%	Financial analysts	0.0%
262	Hotel managers	0.0%	Petroleum and natural gas refining plant operators	0.0%
263	Film, stage, and related directors and producers	0.0%	Information and communications technology operations technicians	0.0%
264	Pet groomers and animal care workers	0.0%	Information and communications technology user support technicians	0.0%
265	Protective services workers not elsewhere classified	0.0%	Other language teachers	0.0%
266	Dieticians and nutritionists	0.0%	Musical instrument makers and tuners	0.0%
267	Other arts teachers	0.0%	Veterinary technicians and assistants	0.0%
268	Research and development managers	0.0%	Photographic products machine operators	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
269	Garden and horticultural labourers	0.0%	Information and communications technology service managers	0.0%
270	Concrete placers, concrete finishers, and related workers	0.0%	Filing and copying clerks	0.0%
271	Petroleum and natural gas refining plant operators	0.0%	Mechanical machinery assemblers	0.0%
272	Cement, stone, and other mineral products machine operators	0.0%	Telecommunications engineering technicians	0.0%
273	Sociologists, anthropologists, and related professionals	0.0%	Earthmoving and related plant operators	0.0%
274	Training and staff development professionals	0.0%	Rubber products machine operators	0.0%
275	Web and multimedia developers	0.0%	Journalists	0.0%
276	Refuse sorters	0.0%	University and higher education teachers	0.0%
277	Sewing, embroidery, and related workers	0.0%	Electronics engineers	0.0%
278	Handicraft workers in wood, basketry, and related materials	0.0%	Special needs teachers	0.0%
279	Travel attendants and travel stewards	0.0%	Aircraft engine mechanics and repairers	0.0%
280	Government tax and excise officials	0.0%	Telecommunications engineers	0.0%
281	Dairy-products makers	0.0%	Personnel and careers professionals	0.0%
282	Aged care services managers	0.0%	Client information workers not elsewhere classified	0.0%
283	Pawnbrokers and money-lenders	0.0%	Mining and metallurgical technicians	0.0%
284	Advertising and public relations managers	0.0%	Medical imaging and therapeutic equipment technicians	0.0%
285	Fibre preparing, spinning, and winding machine operators	0.0%	Clerical support workers not elsewhere classified	0.0%
286	Sports, recreation, and cultural centre managers	0.0%	Judges	0.0%
287	Dispensing opticians	0.0%	Optometrists and ophthalmic opticians	0.0%
288	Blacksmiths, hammersmiths, and forging press workers	0.0%	Life science technicians (excluding medical)	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
289	Vocational education teachers	0.0%	Railway brake, signal, and switch operators	0.0%
290	Photographic products machine operators	0.0%	Craft and related workers not elsewhere classified	0.0%
291	Shop supervisors	0.0%	Business services and administration managers not elsewhere classified	0.0%
292	Manufacturing labourers not elsewhere classified	0.0%	Mixed crop and livestock farm labourers	0.0%
293	Well drillers and borers and related workers	0.0%	Physiotherapists	0.0%
294	Visual artists	0.0%	Survey and market research interviewers	0.0%
295	Personnel clerks	0.0%	Psychologists	0.0%
296	Computer network and systems technicians	0.0%	Religious associate professionals	0.0%
297	Life science technicians (excluding medical)	0.0%	Animal producers not elsewhere classified	0.0%
298	Software developers	0.0%	Ships' deck crews and related workers	0.0%
299	Data entry clerks	0.0%	Statistical, finance, and insurance clerks	0.0%
300	Agricultural and forestry production managers	0.0%	Bank tellers and related clerks	0.0%
301	Craft and related workers not elsewhere classified	0.0%	Health professionals not elsewhere classified	0.0%
302	Production clerks	0.0%	Apiarists and sericulturists	0.0%
303	Lifting truck operators	0.0%	Ships' deck officers and pilots	0.0%
304	Textile, fur, and leather products machine operators not elsewhere classified	0.0%	Aquaculture and fisheries production managers	0.0%
305	Dancers and choreographers	0.0%	Ambulance workers	0.0%
306	Ships' deck officers and pilots	0.0%	Riggers and cable splicers	0.0%
307	Systems administrators	0.0%	Management and organization analysts	0.0%
308	Environmental and occupational health and hygiene professionals	0.0%	Protective services workers not elsewhere classified	0.0%
309	Motorcycle drivers	0.0%	Employment agents and contractors	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
310	Musical instrument makers and tuners	0.0%	Building structure cleaners	0.0%
311	Potters and related workers	0.0%	Electrical and electronic equipment assemblers	0.0%
312	Optometrists and ophthalmic opticians	0.0%	Handicraft workers not elsewhere classified	0.0%
313	Public relations professionals	0.0%	Education managers	0.0%
314	Sign writers, decorative painters, engravers, and etchers	0.0%	Announcers on radio, television, and other media	0.0%
315	Geologists and geophysicists	0.0%	Personnel clerks	0.0%
316	Metal polishers, wheel grinders, and tool sharpeners	0.0%	Legal secretaries	0.0%
317	Travel guides	0.0%	Environmental protection professionals	0.0%
318	Creative and performing artists not elsewhere classified	0.0%	Mathematicians, actuaries, and statisticians	0.0%
319	Metal production process controllers	0.0%	Accounting associate professionals	0.0%
320	Insulation workers	0.0%	Garden and horticultural labourers	0.0%
321	Forestry and related workers	0.0%	Wood treaters	0.0%
322	Religious associate professionals	0.0%	Research and development managers	0.0%
323	Pulp and papermaking plant operators	0.0%	Finance managers	0.0%
324	Statistical, mathematical, and related associate professionals	0.0%	Sweepers and related labourers	0.0%
325	Glass makers, cutters, grinders, and finishers	0.0%	Other music teachers	0.0%
326	Child care services managers	0.0%	Authors and related writers	0.0%
327	Steam engine and boiler operators	0.0%	Mining engineers, metallurgists, and related professionals	0.0%
328	Philosophers, historians, and political scientists	0.0%	Data entry clerks	0.0%
329	Fur and leather preparing machine operators	0.0%	Commissioned armed forces officers	0.0%
330	Glaziers	0.0%	Chemical processing plant controllers	0.0%
331	Bicycle and related repairers	0.0%	Philosophers, historians, and political scientists	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
332	Mechanical machinery assemblers	0.0%	Database designers and administrators	0.0%
333	Domestic cleaners and helpers	0.0%	Health services managers	0.0%
334	Meteorologists	0.0%	Child care services managers	0.0%
335	Clerical support workers not elsewhere classified	0.0%	Sociologists, anthropologists, and related professionals	0.0%
336	Rubber products machine operators	0.0%	Well drillers and borers and related workers	0.0%
337	Tobacco preparers and tobacco products makers	0.0%	Legislators	0.0%
338	Fishery and aquaculture labourers	0.0%	Environmental and occupational health and hygiene professionals	0.0%
339	Inland and coastal waters fishery workers	0.0%	Other cleaning workers	0.0%
340	Underwater divers	0.0%	Environmental engineers	0.0%
341	Mining managers	0.0%	Professional services managers not elsewhere classified	0.0%
342	Computer network professionals	0.0%	Government tax and excise officials	0.0%
343	Sales workers not elsewhere classified	0.0%	Shop supervisors	0.0%
344	Health associate professionals not elsewhere classified	0.0%	Librarians and related information professionals	0.0%
345	Roofers	0.0%	Film, stage, and related directors and producers	0.0%
346	Stationary plant and machine operators not elsewhere classified	0.0%	Training and staff development professionals	0.0%
347	Physiotherapy technicians and assistants	0.0%	Systems administrators	0.0%
348	General office clerks	0.0%	Mixed crop and animal producers	0.0%
349	Clearing and forwarding agents	0.0%	Ships' engineers	0.0%
350	Domestic housekeepers	0.0%	Window cleaners	0.0%
351	Other music teachers	0.0%	Farming, forestry, and fisheries advisers	0.0%
352	Pelt dressers, tanners, and fellmongers	0.0%	Medical secretaries	0.0%
353	Ships' deck crews and related workers	0.0%	Economists	0.0%
354	Plasterers	0.0%	Mining supervisors	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
355	Legal secretaries	0.0%	Pulp and papermaking plant operators	0.0%
356	Veterinary technicians and assistants	0.0%	Metal moulders and coremakers	0.0%
357	Aquaculture workers	0.0%	Actors	0.0%
358	Air traffic controllers	0.0%	Public relations professionals	0.0%
359	Street food salespersons	0.0%	Social welfare managers	0.0%
360	Undertakers and embalmers	0.0%	Process control technicians not elsewhere classified	0.0%
361	Street and related service workers	0.0%	Health associate professionals not elsewhere classified	0.0%
362	Software and applications developers and analysts not elsewhere classified	0.0%	Stationary plant and machine operators not elsewhere classified	0.0%
363	Electrical and electronic equipment assemblers	0.0%	Astrologers, fortune-tellers, and related workers	0.0%
364	Forestry labourers	0.0%	Production clerks	0.0%
365	Locomotive engine drivers	0.0%	Library clerks	0.0%
366	Chemical processing plant controllers	0.0%	Underwater divers	0.0%
367	Forestry technicians	0.0%	Statistical, mathematical, and related associate professionals	0.0%
368	Drivers of animal-drawn vehicles and machinery	0.0%	Trade brokers	0.0%
369	Information technology trainers	0.0%	Visual artists	0.0%
370	Tree and shrub crop growers	0.0%	Travel attendants and travel stewards	0.0%
371	Ships' engineers	0.0%	Dancers and choreographers	0.0%
372	Metal moulders and coremakers	0.0%	Vocational education teachers	0.0%
373	Actors	0.0%	Police officers	0.0%
374	Air traffic safety electronics technicians	0.0%	Software developers	0.0%
375	Building structure cleaners	0.0%	Undertakers and embalmers	0.0%
376	Ambulance workers	0.0%	Regulatory government associate professionals not elsewhere classified	0.0%
377	Customs and border inspectors	0.0%	Typists and word processing operators	0.0%
378	Animal producers not elsewhere classified	0.0%	Web technicians	0.0%

#	ISCO title	Formal workers	ISCO title	Informal workers
379	Coding, proof-reading, and related clerks	0.0%	Computer network and systems technicians	0.0%
380	Other cleaning workers	0.0%	Metal production process controllers	0.0%
381	Deep-sea fishery workers	0.0%	Town and traffic planners	0.0%
382	Government licensing officials	0.0%	Geologists and geophysicists	0.0%
383	Web technicians	0.0%	Forestry technicians	0.0%
384	Government social benefits officials	0.0%	Government licensing officials	0.0%
385	Gallery, museum, and library technicians	0.0%	Software and applications developers and analysts not elsewhere classified	0.0%
386	Mixed crop and livestock farm labourers	0.0%	Government social benefits officials	0.0%
387	Medical secretaries	0.0%	Locomotive engine drivers	0.0%
388	Audiologists and speech therapists	0.0%	Dieticians and nutritionists	0.0%
389	Shelf fillers	0.0%	Non-commissioned armed forces officers	0.0%
390	Policy and planning managers	0.0%	Mining managers	0.0%
391	Town and traffic planners	0.0%		
392	Social welfare managers	0.0%		
393	Wood treaters	0.0%		
394	Astrologers, fortune-tellers, and related workers	0.0%		
395	Apiarists and sericulturists	0.0%		
396	Regulatory government associate professionals not elsewhere classified	0.0%		
397	Trade brokers	0.0%		
398	Typists and word processing operators	0.0%		
399	Mixed crop and animal producers	0.0%		
400	Medical records and health information technicians	0.0%		
401	Odd job persons	0.0%		
402	Window cleaners	0.0%		

Source: Author's calculations based on GEIH information, 2016-2018.

Table H.3. **Occupational distribution of the unemployed in Colombia**

#	ISCO title	Unemployed
1	Sales demonstrators	13.9%
2	Cleaners and helpers in offices, hotels, and other establishments	4.9%
3	Domestic cleaners and helpers	4.4%
4	Building and related electricians	3.2%
5	Waiters	3.1%
6	Security guards	3.1%
7	Stock clerks	2.7%
8	Car, taxi, and van drivers	2.7%
9	Health care assistants	2.0%
10	Accounting and bookkeeping clerks	2.0%
11	Secondary education teachers	2.0%
12	Policy administration professionals	1.7%
13	Kitchen helpers	1.6%
14	Contact centre information clerks	1.6%
15	Cooks	1.6%
16	Cashiers and ticket clerks	1.5%
17	Bricklayers and related workers	1.5%
18	Sewing machine operators	1.4%
19	Child care workers	1.2%
20	Construction supervisors	1.1%
21	House builders	1.0%
22	Senior officials of special-interest organizations	1.0%
23	Real estate agents and property managers	0.9%
24	Mail carriers and sorting clerks	0.8%
25	Heavy truck and lorry drivers	0.7%
26	Painters and related workers	0.7%
27	Restaurant managers	0.7%
28	Commercial sales representatives	0.7%
29	Bakers, pastry-cooks, and confectionery makers	0.6%
30	Welders and flame cutters	0.6%
31	Motor vehicle mechanics and repairers	0.6%
32	Filing and copying clerks	0.6%

#	ISCO title	Unemployed
33	Primary school teachers	0.6%
34	Business services agents not elsewhere classified	0.6%
35	Civil engineers	0.6%
36	Stonemasons, stone cutters, splitters, and carvers	0.6%
37	Shop sales assistants	0.6%
38	Crop farm labourers	0.5%
39	Psychologists	0.5%
40	Administrative and executive secretaries	0.5%
41	Receptionists (general)	0.5%
42	Driving instructors	0.5%
43	Freight handlers	0.5%
44	Human resource managers	0.5%
45	Agricultural and industrial machinery mechanics and repairers	0.4%
46	Graphic and multimedia designers	0.4%
47	Packing, bottling, and labelling machine operators	0.4%
48	Butchers, fishmongers, and related food preparers	0.4%
49	Beauticians and related workers	0.4%
50	Secretaries (general)	0.4%
51	Nursing professionals	0.4%
52	Industrial and production engineers	0.4%
53	Environmental and occupational health inspectors and associates	0.4%
54	Enquiry clerks	0.4%
55	Physical and engineering science technicians not elsewhere classified	0.4%
56	Stall and market salespersons	0.4%
57	Services managers not elsewhere classified	0.3%
58	Civil engineering technicians	0.3%
59	Legal professionals not elsewhere classified	0.3%
60	Lawyers	0.3%
61	Social work and counselling professionals	0.3%
62	Information and communications technology sales professionals	0.3%
63	Systems analysts	0.3%
64	Financial and insurance services branch managers	0.3%
65	Manufacturing supervisors	0.3%

#	ISCO title	Unemployed
66	Bartenders	0.3%
67	Fast food preparers	0.3%
68	Pharmaceutical technicians and assistants	0.3%
69	Building construction labourers	0.3%
70	Building architects	0.3%
71	Shoemakers and related workers	0.3%
72	Personal care workers in health services not elsewhere classified	0.3%
73	Information and communications technology installers and servicers	0.3%
74	Messengers, package deliverers and luggage porters	0.3%
75	Archivists and curators	0.2%
76	Carpenters and joiners	0.2%
77	Crane, hoist, and related plant operators	0.2%
78	Supply, distribution, and related managers	0.2%
79	Survey and market research interviewers	0.2%
80	Hand launderers and pressers	0.2%
81	Buyers	0.2%
82	Statistical, finance, and insurance clerks	0.2%
83	Gardeners, horticultural, and nursery growers	0.2%
84	Retail and wholesale trade managers	0.2%
85	Transport clerks	0.2%
86	Environmental engineers	0.2%
87	Translators, interpreters, and other linguists	0.2%
88	Securities and finance dealers and brokers	0.2%
89	Product and garment designers	0.2%
90	Agricultural technicians	0.2%
91	Mechanical engineers	0.2%
92	Fruit, vegetable, and related preservers	0.2%
93	Miners and quarriers	0.2%
94	Information and communications technology operations technicians	0.2%
95	Bank tellers and related clerks	0.2%
96	Electronics engineers	0.2%
97	Database and network professionals not elsewhere classified	0.2%
98	Paper products machine operators	0.2%

#	ISCO title	Unemployed
99	Sweepers and related labourers	0.2%
100	Hotel receptionists	0.2%
101	Teaching professionals not elsewhere classified	0.2%
102	Food and related products machine operators	0.2%
103	Electrical mechanics and fitters	0.2%
104	Hand packers	0.2%
105	Police inspectors and detectives	0.2%
106	Office supervisors	0.2%
107	Door to door salespersons	0.2%
108	Farming, forestry, and fisheries advisers	0.2%
109	Information and communications technology user support technicians	0.1%
110	Early childhood educators	0.1%
111	Generalist medical practitioners	0.1%
112	Cabinet-makers and related workers	0.1%
113	Biologists, botanists, zoologists, and related professionals	0.1%
114	Chefs	0.1%
115	Construction managers	0.1%
116	Electronics mechanics and servicers	0.1%
117	Specialist medical practitioners	0.1%
118	Dentists	0.1%
119	Electrical engineers	0.1%
120	Engineering professionals not elsewhere classified	0.1%
121	Field crop and vegetable growers	0.1%
122	Journalists	0.1%
123	Home-based personal care workers	0.1%
124	Musicians, singers, and composers	0.1%
125	Telecommunications engineers	0.1%
126	Accountants	0.1%
127	Electrical engineering technicians	0.1%
128	Advertising and marketing professionals	0.1%
129	Armed forces occupations, other ranks	0.1%
130	Manufacturing managers	0.1%
131	Other language teachers	0.1%

#	ISCO title	Unemployed
132	Bus and tram drivers	0.1%
133	Food service counter attendants	0.1%
134	Building caretakers	0.1%
135	Sports coaches, instructors, and officials	0.1%
136	Insurance representatives	0.1%
137	Fitness and recreation instructors and program leaders	0.1%
138	Assemblers not elsewhere classified	0.1%
139	Hairdressers	0.1%
140	Financial and investment advisers	0.1%
141	Teachers' aides	0.1%
142	Financial analysts	0.1%
143	Chemists	0.1%
144	Sales and marketing managers	0.1%
145	University and higher education teachers	0.1%
146	Mechanical engineering technicians	0.1%
147	Mining engineers, metallurgists, and related professionals	0.1%
148	Managing directors and chief executives	0.1%
149	Physiotherapists	0.1%
150	Plastic products machine operators	0.1%
151	Telephone switchboard operators	0.1%
152	Personnel and careers professionals	0.1%
153	Shop keepers	0.1%
154	Livestock and dairy producers	0.1%
155	Economists	0.1%
156	Tailors, dressmakers, furriers, and hatters	0.1%
157	Other artistic and cultural associate professionals	0.1%
158	Conference and event planners	0.1%
159	Veterinarians	0.1%
160	Draughtspersons	0.1%
161	Personal services workers not elsewhere classified	0.1%
162	Mathematicians, actuaries, and statisticians	0.1%
163	Sheet-metal workers	0.1%
164	Information and communications technology service managers	0.1%

#	ISCO title	Unemployed
165	Chemical engineers	0.1%
166	Cartographers and surveyors	0.1%
167	Bookmakers, croupiers, and related gaming workers	0.1%
168	Broadcasting and audio-visual technicians	0.1%
169	Toolmakers and related workers	0.1%
170	Shoemaking and related machine operators	0.1%
171	Sports, recreation, and cultural centre managers	0.1%
172	Medical and pathology laboratory technicians	0.1%
173	Upholsterers and related workers	0.1%
174	Meter readers and vending-machine collectors	0.1%
175	Glass and ceramics plant operators	0.1%
176	Contact centre salespersons	0.1%
177	Management and organization analysts	0.1%
178	Concrete placers, concrete finishers, and related workers	0.1%
179	Sociologists, anthropologists, and related professionals	0.1%
180	Special needs teachers	0.1%
181	Electronics engineering technicians	0.1%
182	Health services managers	0.1%
183	Aquaculture and fisheries production managers	0.1%
184	Electrical line installers and repairers	0.1%
185	Vehicle cleaners	0.1%
186	Aircraft pilots and related associate professionals	0.1%
187	Payroll clerks	0.1%
188	Air conditioning and refrigeration mechanics	0.1%
189	Metal working machine tool setters and operators	0.1%
190	Police officers	0.1%
191	Handicraft workers in textile, leather, and related materials	0.1%
192	Motorcycle drivers	0.1%
193	Garbage and recycling collectors	0.1%
194	Fumigators and other pest and weed controllers	0.0%
195	Laundry machine operators	0.0%
196	Spray painters and varnishers	0.0%
197	Mobile farm and forestry plant operators	0.0%

#	ISCO title	Unemployed
198	Dental assistants and therapists	0.0%
199	Education methods specialists	0.0%
200	Travel guides	0.0%
201	Cleaning and housekeeping supervisors in offices, hotels, and other establishments	0.0%
202	Photographers	0.0%
203	Garment and related pattern-makers and cutters	0.0%
204	Plumbers and pipe fitters	0.0%
205	Elementary workers not elsewhere classified	0.0%
206	Service station attendants	0.0%
207	Sewing, embroidery, and related workers	0.0%
208	Technical and medical sales professionals (excluding ICT)	0.0%
209	Chemical engineering technicians	0.0%
210	Finance managers	0.0%
211	Meteorologists	0.0%
212	Railway brake, signal, and switch operators	0.0%
213	Agricultural and forestry production managers	0.0%
214	Database designers and administrators	0.0%
215	Geologists and geophysicists	0.0%
216	Street vendors (excluding food)	0.0%
217	Poultry producers	0.0%
218	Chemical and physical science technicians	0.0%
219	Film, stage, and related directors and producers	0.0%
220	Building frame and related trades workers not elsewhere classified	0.0%
221	Credit and loans officers	0.0%
222	Travel attendants and travel stewards	0.0%
223	Interior designers and decorators	0.0%
224	Precision-instrument makers and repairers	0.0%
225	Fire-fighters	0.0%
226	Craft and related workers not elsewhere classified	0.0%
227	Jewellery and precious-metal workers	0.0%
228	Librarians and related information professionals	0.0%
229	Sales workers not elsewhere classified	0.0%

#	ISCO title	Unemployed
230	Veterinary technicians and assistants	0.0%
231	Earthmoving and related plant operators	0.0%
232	Philosophers, historians, and political scientists	0.0%
233	Mining and metallurgical technicians	0.0%
234	Announcers on radio, television, and other media	0.0%
235	Companions and valets	0.0%
236	Structural-metal preparers and erectors	0.0%
237	Data entry clerks	0.0%
238	Client information workers not elsewhere classified	0.0%
239	Print finishing and binding workers	0.0%
240	Wood processing plant operators	0.0%
241	Medical imaging and therapeutic equipment technicians	0.0%
242	Handicraft workers in wood, basketry, and related materials	0.0%
243	Medical and dental prosthetic technicians	0.0%
244	Incinerator and water treatment plant operators	0.0%
245	Metal processing plant operators	0.0%
246	Telecommunications engineering technicians	0.0%
247	Debt-collectors and related workers	0.0%
248	Aircraft engine mechanics and repairers	0.0%
249	Pet groomers and animal care workers	0.0%
250	Refuse sorters	0.0%
251	Bleaching, dyeing, and fabric cleaning machine operators	0.0%
252	Health professionals not elsewhere classified	0.0%
253	Chemical products plant and machine operators	0.0%
254	Hotel managers	0.0%
255	Clearing and forwarding agents	0.0%
256	Social work associate professionals	0.0%
257	Mining supervisors	0.0%
258	Livestock farm labourers	0.0%
259	Civil engineering labourers	0.0%
260	Environmental protection professionals	0.0%
261	Floor layers and tile setters	0.0%
262	Research and development managers	0.0%

#	ISCO title	Unemployed
263	Non-commissioned armed forces officers	0.0%
264	Manufacturing labourers not elsewhere classified	0.0%
265	Metal production process controllers	0.0%
266	Production clerks	0.0%
267	Musical instrument makers and tuners	0.0%
268	Potters and related workers	0.0%
269	Life science technicians (excluding medical)	0.0%
270	Shelf fillers	0.0%
271	Other arts teachers	0.0%
272	Computer network professionals	0.0%
273	Printers	0.0%
274	Software developers	0.0%
275	Blacksmiths, hammersmiths, and forging press workers	0.0%
276	Cement, stone, and other mineral products machine operators	0.0%
277	Professional services managers not elsewhere classified	0.0%
278	Street and related service workers	0.0%
279	Accounting associate professionals	0.0%
280	Fishery and aquaculture labourers	0.0%
281	Library clerks	0.0%
282	Computer network and systems technicians	0.0%
283	Pharmacists	0.0%
284	Pre-press technicians	0.0%
285	Business services and administration managers not elsewhere classified	0.0%
286	Vocational education teachers	0.0%
287	Tobacco preparers and tobacco products makers	0.0%
288	Authors and related writers	0.0%
289	Dairy-products makers	0.0%
290	Dieticians and nutritionists	0.0%
291	Well drillers and borers and related workers	0.0%
292	Legal secretaries	0.0%
293	Metal finishing, plating, and coating machine operators	0.0%
294	Environmental and occupational health and hygiene professionals	0.0%
295	Physiotherapy technicians and assistants	0.0%

#	ISCO title	Unemployed
296	Domestic housekeepers	0.0%
297	Pawnbrokers and money-lenders	0.0%
298	Personnel clerks	0.0%
299	Roofers	0.0%
300	Travel consultants and clerks	0.0%
301	Ships' deck crews and related workers	0.0%
302	Plasterers	0.0%
303	Electrical and electronic equipment assemblers	0.0%
304	Legislators	0.0%
305	Tree and shrub crop growers	0.0%
306	Riggers and cable splicers	0.0%
307	Public relations professionals	0.0%
308	Forestry and related workers	0.0%
309	Religious professionals	0.0%
310	Underwater divers	0.0%
311	Process control technicians not elsewhere classified	0.0%
312	Insulation workers	0.0%
313	Photographic products machine operators	0.0%
314	Training and staff development professionals	0.0%
315	Weaving and knitting machine operators	0.0%
316	Glaziers	0.0%
317	Advertising and public relations managers	0.0%
318	Bicycle and related repairers	0.0%
319	Rubber products machine operators	0.0%
320	Clerical support workers not elsewhere classified	0.0%
321	Forestry labourers	0.0%
322	Aquaculture workers	0.0%
323	Education managers	0.0%
324	Air traffic controllers	0.0%
325	Deep-sea fishery workers	0.0%
326	Systems administrators	0.0%
327	Shop supervisors	0.0%
328	Creative and performing artists not elsewhere classified	0.0%

#	ISCO title	Unemployed
329	Forestry technicians	0.0%
330	Chemical processing plant controllers	0.0%
331	Government tax and excise officials	0.0%
332	Visual artists	0.0%
333	Other music teachers	0.0%
334	Glass makers, cutters, grinders, and finishers	0.0%
335	Apiarists and sericulturists	0.0%
336	Mineral and stone processing plant operators	0.0%
337	Town and traffic planners	0.0%
338	Dancers and choreographers	0.0%
339	Aged care services managers	0.0%
340	Sign writers, decorative painters, engravers, and etchers	0.0%
341	Software and applications developers and analysts not elsewhere classified	0.0%
342	Lifting truck operators	0.0%
343	Garden and horticultural labourers	0.0%
344	Ambulance workers	0.0%
345	Fur and leather preparing machine operators	0.0%
346	Street food salespersons	0.0%
347	General office clerks	0.0%
348	Protective services workers not elsewhere classified	0.0%
349	Web and multimedia developers	0.0%
350	Steam engine and boiler operators	0.0%
351	Dispensing opticians	0.0%
352	Transport conductors	0.0%
353	Commissioned armed forces officers	0.0%
354	Ships' deck officers and pilots	0.0%
355	Judges	0.0%
356	Undertakers and embalmers	0.0%
357	Fibre preparing, spinning, and winding machine operators	0.0%
358	Inland and coastal waters fishery workers	0.0%
359	Petroleum and natural gas refining plant operators	0.0%
360	Health associate professionals not elsewhere classified	0.0%

#	ISCO title	Unemployed
361	Optometrists and ophthalmic opticians	0.0%
362	Textile, fur, and leather products machine operators not elsewhere classified	0.0%
363	Employment agents and contractors	0.0%
364	Child care services managers	0.0%
365	Gallery, museum, and library technicians	0.0%
366	Handicraft workers not elsewhere classified	0.0%
367	Mining and quarrying labourers	0.0%
368	Government social benefits officials	0.0%
369	Air traffic safety electronics technicians	0.0%
370	Animal producers not elsewhere classified	0.0%
371	Mining managers	0.0%

Source: Author's calculations based on GEIH information, 2016-2018.

Table H.4. **Informality rate by occupation**

#	ISCO title	Informality rate
1	Domestic cleaners and helpers	99.8%
2	Motorcycle drivers	99.0%
3	Shop keepers	97.3%
4	Tailors, dressmakers, furriers, and hatters	96.7%
5	Street food salespersons	96.6%
6	Stall and market salespersons	95.3%
7	Sewing, embroidery, and related workers	94.1%
8	Drivers of animal-drawn vehicles and machinery	93.6%
9	Potters and related workers	92.3%
10	Clearing and forwarding agents	92.2%
11	Sales workers not elsewhere classified	92.0%
12	Beauticians and related workers	90.7%
13	Handicraft workers in textile, leather, and related materials	90.7%
14	Hairdressers	89.2%
15	Bicycle and related repairers	89.0%
16	Fast food preparers	87.6%
17	Laundry machine operators	87.2%

#	ISCO title	Informality rate
18	Refuse sorters	86.0%
19	Street vendors (excluding food)	84.9%
20	Bricklayers and related workers	83.7%
21	Pawnbrokers and money-lenders	81.7%
22	Sales demonstrators	81.5%
23	Shoemakers and related workers	81.5%
24	Hand launderers and pressers	80.7%
25	Dairy-products makers	80.7%
26	Sheet-metal workers	80.5%
27	Miners and quarriers	80.4%
28	Contact centre salespersons	80.2%
29	Crop farm labourers	79.8%
30	Meteorologists	79.3%
31	Handicraft workers in wood, basketry, and related materials	78.7%
32	Tobacco preparers and tobacco products makers	78.7%
33	Motor vehicle mechanics and repairers	78.6%
34	Home-based personal care workers	78.6%
35	Cabinet-makers and related workers	78.5%
36	Painters and related workers	78.5%
37	Jewellery and precious-metal workers	77.7%
38	Street and related service workers	77.5%
39	Pet groomers and animal care workers	76.9%
40	Inland and coastal waters fishery workers	76.6%
41	Database and network professionals not elsewhere classified	76.5%
42	Forestry and related workers	76.5%
43	Gardeners, horticultural, and nursery growers	76.1%
44	Physiotherapy technicians and assistants	76.0%
45	Food service counter attendants	75.8%
46	Forestry labourers	75.7%
47	Carpenters and joiners	75.4%
48	Construction supervisors	75.4%
49	Commercial sales representatives	75.2%
50	Garbage and recycling collectors	75.1%

#	ISCO title	Informality rate
51	Door to door salespersons	75.0%
52	Stonemasons, stone cutters, splitters, and carvers	75.0%
53	Poultry producers	74.6%
54	Lifting truck operators	74.4%
55	Domestic housekeepers	74.2%
56	Cooks	74.0%
57	Concrete placers, concrete finishers, and related workers	73.0%
58	Upholsterers and related workers	72.6%
59	Fruit, vegetable, and related preservers	72.5%
60	Car, taxi, and van drivers	71.9%
61	Building caretakers	71.8%
62	Bookmakers, croupiers, and related gaming workers	71.6%
63	Credit and loans officers	71.6%
64	Companions and valets	71.4%
65	Garment and related pattern-makers and cutters	71.3%
66	Spray painters and varnishers	71.2%
67	Freight handlers	71.1%
68	Roofers	69.9%
69	Real estate agents and property managers	69.9%
70	Services managers not elsewhere classified	69.7%
71	Steam engine and boiler operators	69.3%
72	Vehicle cleaners	69.1%
73	Agricultural and forestry production managers	69.0%
74	Food and related products machine operators	68.2%
75	Bakers, pastry-cooks, and confectionery makers	68.1%
76	Bartenders	67.9%
77	Fur and leather preparing machine operators	67.6%
78	Fumigators and other pest and weed controllers	66.9%
79	Plasterers	66.8%
80	Sign writers, decorative painters, engravers, and etchers	65.6%
81	Floor layers and tile setters	65.1%
82	House builders	64.9%
83	Child care workers	64.2%

#	ISCO title	Informality rate
84	Metal polishers, wheel grinders, and tool sharpeners	63.8%
85	Glaziers	63.4%
86	Pelt dressers, tanners, and fellmongers	63.3%
87	Sewing machine operators	62.7%
88	Travel guides	62.2%
89	Livestock farm labourers	61.6%
90	Messengers, package deliverers, and luggage porters	61.5%
91	Butchers, fishmongers, and related food preparers	61.1%
92	Shop sales assistants	60.9%
93	Veterinary technicians and assistants	60.8%
94	Printers	59.8%
95	Teachers' aides	59.7%
96	Restaurant managers	59.0%
97	Kitchen helpers	59.0%
98	Hotel managers	58.6%
99	Building and related electricians	58.3%
100	Medical and dental prosthetic technicians	57.4%
101	Heavy truck and lorry drivers	56.3%
102	Blacksmiths, hammersmiths, and forging press workers	56.0%
103	Cement, stone, and other mineral products machine operators	54.5%
104	Other artistic and cultural associate professionals	54.2%
105	Cleaners and helpers in offices, hotels, and other establishments	54.1%
106	Product and garment designers	54.1%
107	Pre-press technicians	53.6%
108	Shoemaking and related machine operators	53.6%
109	Print finishing and binding workers	53.5%
110	Glass makers, cutters, grinders, and finishers	53.2%
111	Manufacturing supervisors	52.1%
112	Aged care services managers	51.6%
113	Insulation workers	51.6%
114	Sports, recreation, and cultural centre managers	51.5%
115	Weaving and knitting machine operators	51.3%
116	Waiters	50.1%

#	ISCO title	Informality rate
117	Assemblers not elsewhere classified	49.6%
118	Manufacturing labourers not elsewhere classified	49.0%
119	Electrical and electronic equipment assemblers	48.9%
120	Textile, fur, and leather products machine operators not elsewhere classified	48.8%
121	Electrical mechanics and fitters	48.5%
122	Conference and event planners	48.0%
123	Building construction labourers	47.9%
124	Retail and wholesale trade managers	47.8%
125	Creative and performing artists not elsewhere classified	47.6%
126	Hotel receptionists	47.3%
127	Other arts teachers	46.9%
128	Precision-instrument makers and repairers	45.8%
129	Ships’ deck crews and related workers	45.6%
130	Transport conductors	45.4%
131	Packing, bottling, and labelling machine operators	45.4%
132	Rubber products machine operators	45.2%
133	Toolmakers and related workers	45.1%
134	Mechanical machinery assemblers	44.9%
135	Wood processing plant operators	44.4%
136	Fibre preparing, spinning, and winding machine operators	44.0%
137	Debt-collectors and related workers	43.3%
138	Chemical products plant and machine operators	42.8%
139	Electronics mechanics and servicers	41.4%
140	Mineral and stone processing plant operators	41.4%
141	Buyers	41.1%
142	Metal finishing, plating, and coating machine operators	41.0%
143	Clerical support workers not elsewhere classified	39.8%
144	Metal processing plant operators	39.8%
145	Elementary workers not elsewhere classified	39.3%
146	Bleaching, dyeing, and fabric cleaning machine operators	39.1%
147	Dispensing opticians	39.1%
148	Translators, interpreters, and other linguists	38.9%
149	Glass and ceramics plant operators	38.6%

#	ISCO title	Informality rate
150	Musical instrument makers and tuners	37.5%
151	Air conditioning and refrigeration mechanics	36.7%
152	Mail carriers and sorting clerks	36.6%
153	Legal secretaries	36.0%
154	Structural-metal preparers and erectors	35.7%
155	Personal care workers in health services not elsewhere classified	35.4%
156	Chemical engineers	35.3%
157	Bus and tram drivers	34.9%
158	Welders and flame cutters	34.8%
159	Advertising and public relations managers	34.8%
160	Biologists, botanists, zoologists, and related professionals	33.1%
161	Interior designers and decorators	32.2%
162	Religious associate professionals	31.8%
163	Web and multimedia developers	31.2%
164	Managing directors and chief executives	31.0%
165	Civil engineering technicians	30.9%
166	Sports coaches, instructors, and officials	30.7%
167	Metal working machine tool setters and operators	30.4%
168	Other music teachers	30.3%
169	Personal services workers not elsewhere classified	29.7%
170	Photographic products machine operators	29.6%
171	Optometrists and ophthalmic opticians	29.4%
172	Fitness and recreation instructors and program leaders	29.0%
173	Electrical line installers and repairers	29.0%
174	Transport clerks	28.8%
175	Pharmacists	28.5%
176	Veterinarians	28.4%
177	Building frame and related trades workers not elsewhere classified	28.2%
178	Chefs	27.8%
179	Technical and medical sales professionals (excluding ICT)	27.7%
180	Civil engineering labourers	27.5%
181	Broadcasting and audio-visual technicians	27.5%
182	Petroleum and natural gas refining plant operators	27.4%

#	ISCO title	Informality rate
183	Incinerator and water treatment plant operators	27.3%
184	Agricultural and industrial machinery mechanics and repairers	26.6%
185	Service station attendants	26.1%
186	Telephone switchboard operators	26.1%
187	Craft and related workers not elsewhere classified	25.7%
188	Cashiers and ticket clerks	25.7%
189	Medical and pathology laboratory technicians	25.6%
190	Life science technicians (excluding medical)	25.2%
191	Ships' deck officers and pilots	24.7%
192	Dentists	23.9%
193	Cleaning and housekeeping supervisors in offices, hotels, and other establishments	23.6%
194	Plumbers and pipe fitters	23.1%
195	Fire-fighters	23.0%
196	Insurance representatives	22.6%
197	Enquiry clerks	22.6%
198	Information and communications technology installers and servicers	22.5%
199	Mechanical engineering technicians	22.2%
200	Civil engineers	22.0%
201	Senior officials of special-interest organizations	22.0%
202	Advertising and marketing professionals	21.1%
203	Photographers	21.1%
204	Armed forces occupations, other ranks	21.0%
205	Aircraft pilots and related associate professionals	21.0%
206	Crane, hoist, and related plant operators	20.8%
207	Social work associate professionals	20.6%
208	Dental assistants and therapists	20.5%
209	Receptionists (general)	20.5%
210	Security guards	20.4%
211	Electrical engineering technicians	20.3%
212	Aircraft engine mechanics and repairers	19.7%
213	Religious professionals	19.4%
214	Electrical engineers	19.2%
215	Travel consultants and clerks	18.7%

#	ISCO title	Informality rate
216	Draughtspersons	18.6%
217	Philosophers, historians, and political scientists	18.4%
218	Paper products machine operators	18.0%
219	Physical and engineering science technicians not elsewhere classified	17.9%
220	Administrative and executive secretaries	17.8%
221	Business services agents not elsewhere classified	17.7%
222	Secretaries (general)	17.4%
223	Manufacturing managers	17.3%
224	Child care services managers	17.1%
225	Hand packers	17.0%
226	Chemists	16.9%
227	Personnel clerks	16.6%
228	Railway brake, signal, and switch operators	16.6%
229	Industrial and production engineers	16.5%
230	Chemical engineering technicians	16.3%
231	Plastic products machine operators	16.1%
232	Graphic and multimedia designers	14.9%
233	Cartographers and surveyors	14.8%
234	Construction managers	14.8%
235	Policy administration professionals	14.3%
236	Aquaculture and fisheries production managers	13.8%
237	Chemical and physical science technicians	13.8%
238	Stock clerks	13.8%
239	Client information workers not elsewhere classified	13.7%
240	Sales and marketing managers	13.7%
241	Data entry clerks	13.2%
242	Business services and administration managers not elsewhere classified	13.1%
243	Agricultural technicians	13.0%
244	Protective services workers not elsewhere classified	13.0%
245	Pharmaceutical technicians and assistants	13.0%
246	Health professionals not elsewhere classified	12.9%
247	Early childhood educators	12.6%
248	Electronics engineering technicians	12.6%

#	ISCO title	Informality rate
249	Education methods specialists	12.5%
250	Financial and investment advisers	12.5%
251	Musicians, singers, and composers	12.2%
252	Riggers and cable splicers	11.8%
253	Environmental and occupational health and hygiene professionals	11.7%
254	Legal professionals not elsewhere classified	11.6%
255	Accounting and bookkeeping clerks	11.4%
256	Announcers on radio, television, and other media	11.3%
257	Garden and horticultural labourers	11.0%
258	Judges	10.8%
259	Employment agents and contractors	10.8%
260	Mechanical engineers	10.6%
261	Telecommunications engineering technicians	10.5%
262	Environmental protection professionals	10.4%
263	Research and development managers	10.2%
264	Meter readers and vending-machine collectors	10.2%
265	Well drillers and borers and related workers	10.2%
266	Mining and metallurgical technicians	10.1%
267	Medical imaging and therapeutic equipment technicians	10.0%
268	Stationary plant and machine operators not elsewhere classified	10.0%
269	Supply, distribution, and related managers	9.8%
270	Payroll clerks	9.7%
271	Engineering professionals not elsewhere classified	9.7%
272	Health associate professionals not elsewhere classified	9.6%
273	Education managers	9.2%
274	Driving instructors	9.0%
275	Information and communications technology sales professionals	8.9%
276	Securities and finance dealers and brokers	8.8%
277	Sociologists, anthropologists, and related professionals	8.8%
278	Archivists and curators	8.7%
279	Building architects	8.7%
280	Systems administrators	8.7%
281	Pulp and papermaking plant operators	8.1%

#	ISCO title	Informality rate
282	Earthmoving and related plant operators	8.0%
283	Environmental and occupational health inspectors and associates	7.9%
284	Shop supervisors	7.8%
285	Social work and counselling professionals	7.8%
286	Information and communications technology operations technicians	7.7%
287	Primary school teachers	7.6%
288	Specialist medical practitioners	7.5%
289	Special needs teachers	7.3%
290	Commissioned armed forces officers	7.2%
291	Government tax and excise officials	7.2%
292	Teaching professionals not elsewhere classified	7.1%
293	Journalists	6.9%
294	Underwater divers	6.5%
295	Personnel and careers professionals	6.3%
296	Lawyers	6.2%
297	Public relations professionals	6.0%
298	Training and staff development professionals	5.9%
299	Accountants	5.9%
300	Police inspectors and detectives	5.8%
301	Health care assistants	5.7%
302	Electronics engineers	5.7%
303	Authors and related writers	5.7%
304	Financial analysts	5.7%
305	Financial and insurance services branch managers	5.7%
306	Generalist medical practitioners	5.6%
307	Film, stage, and related directors and producers	5.5%
308	Mining engineers, metallurgists, and related professionals	5.4%
309	Information and communications technology user support technicians	5.4%
310	Information and communications technology service managers	5.3%
311	Survey and market research interviewers	4.7%
312	Telecommunications engineers	4.4%
313	Statistical, mathematical, and related associate professionals	4.4%
314	Physiotherapists	3.9%

#	ISCO title	Informality rate
315	Accounting associate professionals	3.8%
316	Professional services managers not elsewhere classified	3.8%
317	Environmental engineers	3.7%
318	Nursing professionals	3.7%
319	Production clerks	3.7%
320	Finance managers	3.7%
321	Office supervisors	3.6%
322	Database designers and administrators	3.5%
323	Human resource managers	3.5%
324	Bank tellers and related clerks	3.4%
325	Other language teachers	3.4%
326	Contact centre information clerks	3.2%
327	Systems analysts	3.1%
328	Process control technicians not elsewhere classified	3.0%
329	Management and organization analysts	2.9%
330	Economists	2.8%
331	Statistical, finance, and insurance clerks	2.8%
332	Health services managers	2.7%
333	Secondary education teachers	2.7%
334	Filing and copying clerks	2.6%
335	Visual artists	2.6%
336	University and higher education teachers	2.5%
337	Sweepers and related labourers	2.4%
338	Software developers	2.1%
339	Vocational education teachers	2.0%
340	Legislators	1.9%
341	Travel attendants and travel stewards	1.9%
342	Mining supervisors	1.8%
343	Metal production process controllers	1.6%
344	Psychologists	1.5%
345	Mathematicians, actuaries, and statisticians	1.2%
346	Computer network and systems technicians	1.2%
347	Geologists and geophysicists	0.9%

#	ISCO title	Informality rate
348	Dieticians and nutritionists	0.3%
349	Computer network professionals	0.0%

Source: Author's calculations based on GEIH information, 2016-2018.

Table H.5. **Unemployment rate by occupation**

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
1	Environmental engineers	36.7%	29.3
2	Geologists and geophysicists	26.1%	31.7
3	Sociologists, anthropologists, and related professionals	25.4%	24.8
4	Economists	22.7%	46.3
5	Philosophers, historians, and political scientists	22.7%	40.3
6	Survey and market research interviewers	22.5%	21.0
7	Contact centre information clerks	22.1%	18.1
8	Filing and copying clerks	21.8%	25.9
9	Veterinary technicians and assistants	21.6%	10.8
10	Environmental and occupational health inspectors and associates	20.7%	27.9
11	Enquiry clerks	20.0%	27.9
12	Mining engineers, metallurgists, and related professionals	19.9%	33.1
13	Receptionists (general)	19.2%	26.1
14	Stock clerks	18.8%	18.6
15	Mechanical engineers	18.7%	25.9
16	Sports, recreation, and cultural centre managers	18.5%	12.9
17	Business services agents not elsewhere classified	18.4%	20.8
18	Social work and counselling professionals	17.9%	29.3
19	Information and communications technology operations technicians	17.5%	24.9
20	Psychologists	17.1%	29.4
21	Electronics engineers	16.9%	38.1
22	Accounting and bookkeeping clerks	16.8%	28.4
23	Travel attendants and travel stewards	16.8%	31.7

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
24	Information and communications technology sales professionals	16.7%	20.4
25	Draughtspersons	16.7%	28.8
26	Sweepers and related labourers	16.6%	23.6
27	Waiters	16.4%	16.9
28	Health care assistants	16.2%	25.0
29	Restaurant managers	16.2%	16.3
30	Agricultural technicians	16.1%	28.9
31	Travel guides	16.0%	24.4
32	Chemical engineers	15.9%	27.4
33	Hand packers	15.7%	28.7
34	Building construction labourers	15.6%	13.4
35	Craft and related workers not elsewhere classified	15.2%	12.0
36	Driving instructors	15.1%	26.6
37	Pharmaceutical technicians and assistants	15.0%	21.9
38	Data entry clerks	14.7%	29.2
39	Bank tellers and related clerks	14.4%	24.2
40	Cashiers and ticket clerks	14.4%	22.5
41	Statistical, finance, and insurance clerks	14.3%	27.1
42	Metal production process controllers	14.3%	20.5
43	Journalists	13.9%	28.7
44	Film, stage, and related directors and producers	13.7%	35.2
45	Personal care workers in health services not elsewhere classified	13.0%	21.8
46	Stonemasons, stone cutters, splitters, and carvers	13.0%	11.8
47	Legal secretaries	13.0%	17.6
48	Aircraft pilots and related associate professionals	12.9%	33.4
49	Hotel receptionists	12.9%	25.9
50	Railway brake, signal, and switch operators	12.7%	25.1
51	Building and related electricians	12.7%	15.9
52	Senior officials of special-interest organizations	12.6%	32.6
53	Building architects	12.4%	27.4
54	Civil engineering technicians	12.4%	19.4

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
55	Civil engineers	12.3%	23.4
56	Electrical and electronic equipment assemblers	12.1%	16.0
57	Cartographers and surveyors	11.8%	23.6
58	Electrical line installers and repairers	11.7%	16.0
59	Packing, bottling, and labelling machine operators	11.5%	18.2
60	Industrial and production engineers	11.4%	30.5
61	Engineering professionals not elsewhere classified	11.3%	27.0
62	Dental assistants and therapists	11.3%	30.0
63	Security guards	11.2%	23.6
64	Graphic and multimedia designers	11.2%	28.9
65	Kitchen helpers	11.2%	20.4
66	Human resource managers	11.1%	26.2
67	Early childhood educators	11.0%	21.6
68	Information and communications technology user support technicians	11.0%	23.5
69	Chemists	11.0%	24.8
70	Production clerks	10.9%	34.1
71	Cleaners and helpers in offices, hotels, and other establishments	10.8%	24.3
72	Real estate agents and property managers	10.5%	24.3
73	Cleaning and housekeeping supervisors in offices, hotels, and other establishments	10.4%	27.5
74	Legal professionals not elsewhere classified	10.4%	27.1
75	Manufacturing supervisors	10.3%	21.4
76	Telephone switchboard operators	10.3%	17.8
77	Bartenders	10.2%	19.2
78	Meteorologists	10.1%	8.5
79	Announcers on radio, television, and other media	10.1%	23.4
80	Telecommunications engineers	10.0%	21.5
81	Nursing professionals	9.9%	25.5
82	Conference and event planners	9.9%	24.5
83	Secretaries (general)	9.8%	27.7
84	House builders	9.8%	11.1

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
85	Domestic cleaners and helpers	9.8%	18.8
86	Personnel and careers professionals	9.7%	20.0
87	Underwater divers	9.7%	4.0
88	Ships' deck crews and related workers	9.6%	8.3
89	Financial and investment advisers	9.6%	29.0
90	Product and garment designers	9.5%	21.8
91	Administrative and executive secretaries	9.3%	28.5
92	Sales demonstrators	9.3%	20.1
93	Securities and finance dealers and brokers	9.1%	27.7
94	Systems analysts	9.0%	29.5
95	Chefs	8.9%	27.6
96	Health services managers	8.9%	20.2
97	Software developers	8.8%	24.6
98	Armed forces occupations, other ranks	8.7%	13.1
99	Concrete placers, concrete finishers, and related workers	8.5%	12.6
100	Veterinarians	8.5%	31.8
101	Agricultural and forestry production managers	8.5%	40.8
102	Translators, interpreters, and other linguists	8.5%	21.1
103	Personal services workers not elsewhere classified	8.4%	19.3
104	Physical and engineering science technicians not elsewhere classified	8.4%	18.2
105	Special needs teachers	8.3%	25.6
106	Construction supervisors	8.3%	15.2
107	Street and related service workers	8.2%	23.6
108	Computer network and systems technicians	8.2%	43.1
109	Bricklayers and related workers	8.1%	12.9
110	Sports coaches, instructors, and officials	8.1%	26.6
111	Biologists, botanists, zoologists, and related professionals	8.1%	31.7
112	Musical instrument makers and tuners	8.0%	34.3
113	Physiotherapists	7.9%	27.5
114	Construction managers	7.8%	14.1

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
115	Buyers	7.8%	21.6
116	Interior designers and decorators	7.7%	21.4
117	Secondary education teachers	7.7%	27.6
118	Agricultural and industrial machinery mechanics and repairers	7.5%	23.2
119	Life science technicians (excluding medical)	7.5%	26.4
120	Paper products machine operators	7.5%	20.7
121	Electronics mechanics and servicers	7.3%	17.9
122	Environmental and occupational health and hygiene professionals	7.3%	42.3
123	Managing directors and chief executives	7.3%	22.4
124	Painters and related workers	7.2%	15.8
125	Chemical engineering technicians	7.2%	18.8
126	Aircraft engine mechanics and repairers	7.1%	18.6
127	Financial and insurance services branch managers	7.1%	23.3
128	Financial analysts	7.1%	21.8
129	Electrical engineers	7.1%	25.0
130	Research and development managers	7.0%	29.0
131	Electrical engineering technicians	7.0%	25.3
132	Mechanical engineering technicians	7.0%	20.1
133	Information and communications technology installers and servicers	7.0%	19.5
134	Database designers and administrators	6.9%	28.5
135	Hand launderers and pressers	6.9%	21.1
136	Fitness and recreation instructors and program leaders	6.9%	23.5
137	Advertising and marketing professionals	6.9%	26.1
138	Vocational education teachers	6.8%	15.9
139	Environmental protection professionals	6.8%	60.8
140	Crane, hoist, and related plant operators	6.8%	14.5
141	Service station attendants	6.8%	19.3
142	Primary school teachers	6.7%	30.6
143	Payroll clerks	6.7%	39.1
144	Mail carriers and sorting clerks	6.6%	23.3

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
145	Client information workers not elsewhere classified	6.6%	16.3
146	Bakers, pastry-cooks, and confectionery makers	6.5%	17.5
147	Sewing machine operators	6.5%	19.6
148	Shop sales assistants	6.5%	19.6
149	Teaching professionals not elsewhere classified	6.4%	24.0
150	Manufacturing managers	6.3%	14.9
151	Electronics engineering technicians	6.3%	21.8
152	Well drillers and borers and related workers	6.3%	38.4
153	Health professionals not elsewhere classified	6.3%	30.4
154	Finance managers	6.2%	28.0
155	Teachers' aides	6.2%	24.9
156	Information and communications technology service managers	6.2%	22.0
157	Public relations professionals	6.0%	15.7
158	Plasterers	6.0%	20.1
159	Building frame and related trades workers not elsewhere classified	6.0%	23.7
160	Supply, distribution, and related managers	6.0%	20.2
161	Heavy truck and lorry drivers	5.9%	23.7
162	Vehicle cleaners	5.9%	17.1
163	Mining and metallurgical technicians	5.8%	26.3
164	Cooks	5.7%	21.4
165	Dentists	5.7%	32.5
166	Management and organization analysts	5.7%	26.1
167	Other language teachers	5.6%	19.4
168	Lawyers	5.6%	31.2
169	Database and network professionals not elsewhere classified	5.6%	20.6
170	Manufacturing labourers not elsewhere classified	5.5%	22.3
171	Personnel clerks	5.5%	27.7
172	Home-based personal care workers	5.4%	26.0
173	Wood processing plant operators	5.4%	9.4
174	Child care workers	5.4%	22.0

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
175	Library clerks	5.3%	4.3
176	Miners and quarriers	5.3%	22.8
177	Butchers, fishmongers, and related food preparers	5.2%	18.9
178	Fruit, vegetable, and related preservers	5.2%	16.7
179	Bus and tram drivers	5.2%	20.3
180	Specialist medical practitioners	5.1%	29.2
181	Welders and flame cutters	5.0%	20.0
182	Librarians and related information professionals	5.0%	42.0
183	Dieticians and nutritionists	5.0%	9.8
184	Medical imaging and therapeutic equipment technicians	5.0%	16.2
185	Clerical support workers not elsewhere classified	5.0%	4.0
186	Accountants	5.0%	29.8
187	Education methods specialists	4.9%	20.1
188	Fast food preparers	4.9%	15.3
189	Domestic housekeepers	4.9%	24.5
190	Roofers	4.9%	8.2
191	Medical and pathology laboratory technicians	4.8%	23.5
192	Elementary workers not elsewhere classified	4.8%	25.8
193	Sales and marketing managers	4.8%	22.8
194	Assemblers not elsewhere classified	4.8%	15.2
195	Non-commissioned armed forces officers	4.8%	10.8
196	Rubber products machine operators	4.8%	12.0
197	Shoemakers and related workers	4.8%	13.4
198	Upholsterers and related workers	4.7%	19.7
199	Metal working machine tool setters and operators	4.6%	15.4
200	Musicians, singers, and composers	4.6%	19.3
201	Car, taxi, and van drivers	4.6%	24.2
202	Plumbers and pipe fitters	4.6%	21.2
203	Insurance representatives	4.6%	26.7
204	Forestry labourers	4.5%	4.0
205	Motor vehicle mechanics and repairers	4.5%	17.7

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
206	Technical and medical sales professionals (excluding ICT)	4.5%	15.7
207	Meter readers and vending-machine collectors	4.4%	24.8
208	Physiotherapy technicians and assistants	4.4%	24.7
209	Broadcasting and audio-visual technicians	4.4%	18.2
210	University and higher education teachers	4.4%	26.1
211	Carpenters and joiners	4.3%	12.7
212	Fire-fighters	4.3%	18.2
213	Precision-instrument makers and repairers	4.3%	28.5
214	Mining supervisors	4.2%	18.0
215	Building caretakers	4.2%	20.9
216	Civil engineering labourers	4.1%	9.3
217	Transport clerks	4.1%	23.8
218	Freight handlers	4.1%	17.6
219	Structural-metal preparers and erectors	4.1%	8.5
220	Sales workers not elsewhere classified	3.9%	15.1
221	Earthmoving and related plant operators	3.9%	43.0
222	Chemical products plant and machine operators	3.9%	33.6
223	Sheet-metal workers	3.8%	16.6
224	Telecommunications engineering technicians	3.8%	20.5
225	Gardeners, horticultural, and nursery growers	3.8%	17.3
226	Crop farm labourers	3.8%	14.0
227	Police officers	3.8%	17.5
228	Plastic products machine operators	3.7%	21.6
229	Mathematicians, actuaries, and statisticians	3.7%	24.9
230	Poultry producers	3.7%	10.2
231	Office supervisors	3.7%	20.9
232	Electrical mechanics and fitters	3.7%	23.0
233	Hotel managers	3.7%	26.6
234	Blacksmiths, hammersmiths, and forging press workers	3.7%	19.0
235	Professional services managers not elsewhere classified	3.6%	27.3
236	Systems administrators	3.6%	12.0

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
237	Social work associate professionals	3.5%	30.1
238	Air conditioning and refrigeration mechanics	3.5%	18.1
239	Training and staff development professionals	3.5%	7.0
240	Messengers, package deliverers, and luggage porters	3.5%	20.3
241	Toolmakers and related workers	3.5%	16.1
242	Field crop and vegetable growers	3.5%	14.1
243	Insulation workers	3.5%	10.3
244	Food and related products machine operators	3.5%	18.4
245	Garment and related pattern-makers and cutters	3.5%	16.6
246	Other arts teachers	3.5%	8.1
247	Shoemaking and related machine operators	3.4%	15.5
248	Photographers	3.3%	20.0
249	Companions and valets	3.2%	17.3
250	Retail and wholesale trade managers	3.2%	23.4
251	Tobacco preparers and tobacco products makers	3.2%	9.3
252	Generalist medical practitioners	3.2%	18.9
253	Cement, stone, and other mineral products machine operators	3.2%	22.0
254	Police inspectors and detectives	3.2%	20.1
255	Business services and administration managers not elsewhere classified	3.2%	20.7
256	Photographic products machine operators	3.1%	8.7
257	Commercial sales representatives	3.1%	21.1
258	Incinerator and water treatment plant operators	3.1%	10.1
259	Process control technicians not elsewhere classified	3.1%	15.6
260	Other artistic and cultural associate professionals	3.1%	24.5
261	Glass and ceramics plant operators	3.1%	9.6
262	Medical and dental prosthetic technicians	3.1%	11.5
263	Chemical and physical science technicians	3.1%	27.0
264	Glaziers	3.1%	1.0
265	Print finishing and binding workers	3.0%	19.7
266	Handicraft workers in wood, basketry, and related materials	3.0%	21.2

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
267	Clearing and forwarding agents	2.9%	20.1
268	Laundry machine operators	2.9%	12.6
269	Beauticians and related workers	2.9%	19.0
270	Shop supervisors	2.8%	39.4
271	Authors and related writers	2.8%	19.8
272	Bookmakers, croupiers, and related gaming workers	2.7%	15.3
273	Spray painters and varnishers	2.7%	10.5
274	Metal processing plant operators	2.7%	6.0
275	Riggers and cable splicers	2.6%	44.0
276	Advertising and public relations managers	2.6%	46.4
277	Garbage and recycling collectors	2.6%	18.9
278	Metal finishing, plating, and coating machine operators	2.5%	8.8
279	Credit and loans officers	2.5%	23.7
280	Pharmacists	2.5%	23.9
281	Floor layers and tile setters	2.5%	12.0
282	Fumigators and other pest and weed controllers	2.5%	15.4
283	Street vendors (excluding food)	2.4%	17.6
284	Food service counter attendants	2.4%	17.7
285	Cabinet-makers and related workers	2.3%	17.4
286	Creative and performing artists not elsewhere classified	2.3%	17.0
287	Pet groomers and animal care workers	2.2%	25.7
288	Fishery and aquaculture labourers	2.1%	13.7
289	Bleaching, dyeing, and fabric cleaning machine operators	2.1%	21.1
290	Jewellery and precious-metal workers	2.0%	12.6
291	Livestock farm labourers	2.0%	17.4
292	Services managers not elsewhere classified	2.0%	22.9
293	Printers	2.0%	9.3
294	Door to door salespersons	2.0%	20.4
295	Hairdressers	1.9%	19.4
296	Accounting associate professionals	1.9%	17.8
297	Contact centre salespersons	1.9%	19.0

#	ISCO title	Unemployment rate	Duration of unemployment (weeks)
298	Travel consultants and clerks	1.8%	18.0
299	Refuse sorters	1.6%	3.3
300	Education managers	1.6%	36.3
301	Debt-collectors and related workers	1.5%	13.1
302	Sewing, embroidery, and related workers	1.4%	24.2
303	Weaving and knitting machine operators	1.4%	26.0
304	Stall and market salespersons	1.4%	20.7
305	Dairy-products makers	1.3%	15.2
306	Pawnbrokers and money-lenders	1.1%	6.0
307	Handicraft workers in textile, leather, and related materials	1.1%	24.3
308	Potters and related workers	1.0%	8.5
309	Tailors, dressmakers, furriers, and hatters	1.0%	23.6
310	Legislators	0.9%	16.3
311	Bicycle and related repairers	0.9%	13.7
312	Shop keepers	0.7%	16.9
313	Motorcycle drivers	0.5%	8.6

Source: Author's calculations based on GEIH information, 2016-2018.

Table H.6. **Occupations with positive employment growth, 2010-2018**

ISCO-08 code	ISCO title	Skill level
3313	Accounting associate professionals	High
7127	Air conditioning and refrigeration mechanics	Medium
7232	Aircraft engine mechanics and repairers	Medium
6221	Aquaculture workers	Low
2621	Archivists and curators	High
5142	Beauticians and related workers	Low
7234	Bicycle and related repairers	Medium
8331	Bus and tram drivers	Medium
7511	Butchers, fishmongers, and related food preparers	Medium
8322	Car, taxi, and van drivers	Medium
3116	Chemical engineering technicians	High

ISCO-08 code	ISCO title	Skill level
9112	Cleaners and helpers in offices, hotels, and other establishments	Low
3331	Clearing and forwarding agents	High
4229	Client information workers not elsewhere classified	Medium
3322	Commercial sales representatives	High
2523	Computer network professionals	High
7114	Concrete placers, concrete finishers, and related workers	Medium
7549	Craft and related workers not elsewhere classified	Medium
7513	Dairy-products makers	Medium
2521	Database designers and administrators	High
3251	Dental assistants and therapists	High
5152	Domestic housekeepers	Low
3118	Draughtspersons	High
8342	Earthmoving and related plant operators	Medium
1345	Education managers	High
7413	Electrical line installers and repairers	Medium
3114	Electronics engineering technicians	High
2152	Electronics engineers	High
2263	Environmental and occupational health and hygiene professionals	High
3257	Environmental and occupational health inspectors and associates	High
9411	Fast food preparers	Low
8151	Fibre preparing, spinning, and winding machine operators	Medium
1346	Financial and insurance services branch managers	High
7122	Floor layers and tile setters	Medium
8155	Fur and leather preparing machine operators	Medium
6113	Gardeners, horticultural, and nursery growers	Low
7532	Garment and related pattern-makers and cutters	Medium
7315	Glass makers, cutters, grinders, and finishers	Medium
7125	Glaziers	Medium
5141	Hairdressers	Low
9121	Hand launderers and pressers	Low
3259	Health associate professionals not elsewhere classified	High
1342	Health services managers	High
1330	Information and communications technology service managers	High
3512	Information and communications technology user support technicians	High

ISCO-08 code	ISCO title	Skill level
6222	Inland and coastal waters fishery workers	Low
7124	Insulation workers	Medium
3321	Insurance representatives	High
7313	Jewellery and precious-metal workers	Medium
3141	Life science technicians (excluding medical)	High
2421	Management and organization analysts	High
1120	Managing directors and chief executives	High
1321	Manufacturing managers	High
3122	Manufacturing supervisors	High
2120	Mathematicians, actuaries, and statisticians	High
3115	Mechanical engineering technicians	High
8211	Mechanical machinery assemblers	Medium
3214	Medical and dental prosthetic technicians	High
9623	Meter readers and vending-machine collectors	Low
8111	Miners and quarriers	Medium
3117	Mining and metallurgical technicians	High
2146	Mining engineers, metallurgists, and related professionals	High
7231	Motor vehicle mechanics and repairers	Medium
8321	Motorcycle drivers	Medium
2221	Nursing professionals	High
2267	Optometrists and ophthalmic opticians	High
3435	Other artistic and cultural associate professionals	High
8183	Packing, bottling, and labelling machine operators	Medium
5329	Personal care workers in health services not elsewhere classified	Low
2423	Personnel and careers professionals	High
5164	Pet groomers and animal care workers	Low
3213	Pharmaceutical technicians and assistants	High
2262	Pharmacists	High
2633	Philosophers, historians, and political scientists	High
8132	Photographic products machine operators	Medium
2264	Physiotherapists	High
3255	Physiotherapy technicians and assistants	High
7123	Plasterers	Medium
3139	Process control technicians not elsewhere classified	High

ISCO-08 code	ISCO title	Skill level
4322	Production clerks	Medium
5419	Protective services workers not elsewhere classified	Low
2634	Psychologists	High
8171	Pulp and papermaking plant operators	Medium
3334	Real estate agents and property managers	High
4226	Receptionists (general)	Medium
9612	Refuse sorters	Low
1223	Research and development managers	High
7121	Roofers	Medium
1221	Sales and marketing managers	High
7533	Sewing, embroidery, and related workers	Medium
7213	Sheet-metal workers	Medium
5221	Shop keepers	Low
2632	Sociologists, anthropologists, and related professionals	High
2352	Special needs teachers	High
3422	Sports coaches, instructors, and officials	High
4312	Statistical, finance, and insurance clerks	Medium
3314	Statistical, mathematical, and related associate professionals	High
8182	Steam engine and boiler operators	Medium
4321	Stock clerks	Medium
9520	Street vendors (excluding food)	Low
1324	Supply, distribution, and related managers	High
7531	Tailors, dressmakers, furriers, and hatters	Medium
5312	Teachers' aides	Low
2153	Telecommunications engineers	High
8159	Textile, fur, and leather products machine operators not elsewhere classified	Medium
7222	Toolmakers and related workers	Medium
2424	Training and staff development professionals	High
4221	Travel consultants and clerks	Medium
5113	Travel guides	Low
7541	Underwater divers	Medium
7534	Upholsterers and related workers	Medium
2250	Veterinarians	High

ISCO-08 code	ISCO title	Skill level
3240	Veterinary technicians and assistants	High
2320	Vocational education teachers	High
2513	Web and multimedia developers	High
8113	Well drillers and borers and related workers	Medium

Source: Author's calculations based on GEIH information, 2016-2018.

Table H.7. **Occupations with positive real wage trend, 2010-2018**

ISCO-08 code	ISCO title	Skill level
3343	Administrative and executive secretaries	High
1222	Advertising and public relations managers	High
1343	Aged care services managers	High
7233	Agricultural and industrial machinery mechanics and repairers	Medium
1311	Agricultural and forestry production managers	High
7127	Air conditioning and refrigeration mechanics	Medium
3154	Air traffic controllers	High
3155	Air traffic safety electronics technicians	High
6221	Aquaculture workers	Low
2621	Archivists and curators	High
8219	Assemblers not elsewhere classified	Medium
5161	Astrologers, fortune-tellers, and related workers	Low
2266	Audiologists and speech therapists	High
7512	Bakers, pastry-cooks, and confectionery makers	Medium
4211	Bank tellers and related clerks	Medium
5132	Bartenders	Low
7234	Bicycle and related repairers	Medium
2131	Biologists, botanists, zoologists, and related professionals	High
7221	Blacksmiths, hammersmiths, and forging press workers	Medium
7112	Bricklayers and related workers	Medium
7411	Building and related electricians	Medium
5153	Building caretakers	Low
9313	Building construction labourers	Low
7119	Building frame and related trades workers not elsewhere classified	Medium

ISCO-08 code	ISCO title	Skill level
3339	Business services agents not elsewhere classified	High
8331	Bus and tram drivers	Medium
7522	Cabinet-makers and related workers	Medium
7115	Carpenters and joiners	Medium
8114	Cement, stone, and other mineral products machine operators	Medium
3111	Chemical and physical science technicians	High
8131	Chemical products plant and machine operators	Medium
2113	Chemists	High
5311	Child care workers	Low
3112	Civil engineering technicians	High
9112	Cleaners and helpers in offices, hotels, and other establishments	Low
4419	Clerical support workers not elsewhere classified	Medium
4229	Client information workers not elsewhere classified	Medium
4413	Coding, proof-reading, and related clerks	Medium
3513	Computer network and systems technicians	High
2523	Computer network professionals	High
7114	Concrete placers, concrete finishers, and related workers	Medium
3123	Construction supervisors	High
4222	Contact centre information clerks	Medium
5244	Contact centre salespersons	Low
7549	Craft and related workers not elsewhere classified	Medium
8343	Crane, hoist, and related plant operators	Medium
9211	Crop farm labourers	Low
3351	Customs and border inspectors	High
7513	Dairy-products makers	Medium
4132	Data entry clerks	Medium
4214	Debt-collectors and related workers	Medium
2265	Dieticians and nutritionists	High
3254	Dispensing opticians	High
9111	Domestic cleaners and helpers	Low
5152	Domestic housekeepers	Low
3118	Draughtspersons	High
8342	Earthmoving and related plant operators	Medium

ISCO-08 code	ISCO title	Skill level
1345	Education managers	High
2152	Electronics engineers	High
3114	Electronics engineering technicians	High
7412	Electrical mechanics and fitters	Medium
9629	Elementary workers not elsewhere classified	Low
3333	Employment agents and contractors	High
2149	Engineering professionals not elsewhere classified	High
2143	Environmental engineers	High
9411	Fast food preparers	Low
8151	Fibre preparing, spinning, and winding machine operators	Medium
6111	Field crop and vegetable growers	Low
4415	Filing and copying clerks	Medium
2654	Film, stage, and related directors and producers	High
5411	Fire-fighters	Low
3423	Fitness and recreation instructors and program leaders	High
8160	Food and related products machine operators	Medium
5246	Food service counter attendants	Low
9215	Forestry labourers	Low
9333	Freight handlers	Low
7514	Fruit, vegetable, and related preservers	Medium
7544	Fumigators and other pest and weed controllers	Medium
8155	Fur and leather preparing machine operators	Medium
9611	Garbage and recycling collectors	Low
6113	Gardeners, horticultural, and nursery growers	Low
7532	Garment and related pattern-makers and cutters	Medium
4110	General office clerks	Medium
2114	Geologists and geophysicists	High
8181	Glass and ceramics plant operators	Medium
7315	Glass makers, cutters, grinders, and finishers	Medium
7125	Glaziers	Medium
3353	Government social benefits officials	High
2166	Graphic and multimedia designers	High
5141	Hairdressers	Low

ISCO-08 code	ISCO title	Skill level
9121	Hand launderers and pressers	Low
9321	Hand packers	Low
7317	Handicraft workers in wood, basketry, and related materials	Medium
1342	Health services managers	High
8332	Heavy truck and lorry drivers	Medium
5322	Home-based personal care workers	Low
7111	House builders	Medium
2141	Industrial and production engineers	High
3511	Information and communications technology operations technicians	High
1330	Information and communications technology service managers	High
2356	Information technology trainers	High
6222	Inland and coastal waters fishery workers	Low
2642	Journalists	High
2612	Judges	High
9412	Kitchen helpers	Low
8157	Laundry machine operators	Medium
2611	Lawyers	High
2619	Legal professionals not elsewhere classified	High
2622	Librarians and related information professionals	High
8344	Lifting truck operators	Medium
4412	Mail carriers and sorting clerks	Medium
9329	Manufacturing labourers not elsewhere classified	Low
2120	Mathematicians, actuaries, and statisticians	High
3115	Mechanical engineering technicians	High
8211	Mechanical machinery assemblers	Medium
3211	Medical imaging and therapeutic equipment technicians	High
3344	Medical secretaries	High
9621	Messengers, package deliverers, and luggage porters	Low
8122	Metal finishing, plating, and coating machine operators	Medium
7211	Metal moulders and coremakers	Medium
7224	Metal polishers, wheel grinders, and tool sharpeners	Medium
7223	Metal working machine tool setters and operators	Medium
8112	Mineral and stone processing plant operators	Medium

ISCO-08 code	ISCO title	Skill level
8111	Miners and quarriers	Medium
2146	Mining engineers, metallurgists, and related professionals	High
3121	Mining supervisors	High
6130	Mixed crop and animal producers	Low
9213	Mixed crop and livestock farm labourers	Low
8341	Mobile farm and forestry plant operators	Medium
8321	Motorcycle drivers	Medium
7231	Motor vehicle mechanics and repairers	Medium
2652	Musicians, singers, and composers	High
9622	Odd job persons	Low
3341	Office supervisors	High
3435	Other artistic and cultural associate professionals	High
2353	Other language teachers	High
2354	Other music teachers	High
8183	Packing, bottling, and labelling machine operators	Medium
8143	Paper products machine operators	Medium
4313	Payroll clerks	Medium
7535	Pelt dressers, tanners, and fellmongers	Medium
5329	Personal care workers in health services not elsewhere classified	Low
5169	Personal services workers not elsewhere classified	Low
2423	Personnel and careers professionals	High
4416	Personnel clerks	Medium
5164	Pet groomers and animal care workers	Low
3134	Petroleum and natural gas refining plant operators	High
3213	Pharmaceutical technicians and assistants	High
2262	Pharmacists	High
3431	Photographers	High
2264	Physiotherapists	High
7123	Plasterers	Medium
8142	Plastic products machine operators	Medium
7126	Plumbers and pipe fitters	Medium
1213	Policy and planning managers	High
3355	Police inspectors and detectives	High

ISCO-08 code	ISCO title	Skill level
7314	Potters and related workers	Medium
6122	Poultry producers	Low
2341	Primary school teachers	High
7322	Printers	Medium
7323	Print finishing and binding workers	Medium
3139	Process control technicians not elsewhere classified	High
5419	Protective services workers not elsewhere classified	Low
2432	Public relations professionals	High
8171	Pulp and papermaking plant operators	Medium
8312	Railway brake, signal, and switch operators	Medium
4226	Receptionists (general)	Medium
1412	Restaurant managers	High
8141	Rubber products machine operators	Medium
5242	Sales demonstrators	Low
5249	Sales workers not elsewhere classified	Low
2330	Secondary education teachers	High
4120	Secretaries (general)	Medium
5414	Security guards	Low
1114	Senior officials of special-interest organizations	High
1439	Services managers not elsewhere classified	High
5245	Service station attendants	Low
7533	Sewing, embroidery, and related workers	Medium
8153	Sewing machine operators	Medium
7213	Sheet-metal workers	Medium
3151	Ships' engineers	High
7536	Shoemakers and related workers	Medium
8156	Shoemaking and related machine operators	Medium
7316	Sign writers, decorative painters, engravers, and etchers	Medium
2635	Social work and counselling professionals	High
2632	Sociologists, anthropologists, and related professionals	High
2352	Special needs teachers	High
3422	Sports coaches, instructors, and officials	High
3314	Statistical, mathematical, and related associate professionals	High

ISCO-08 code	ISCO title	Skill level
4312	Statistical, finance, and insurance clerks	Medium
4321	Stock clerks	Medium
7113	Stonemasons, stone cutters, splitters, and carvers	Medium
9520	Street vendors (excluding food)	Low
7214	Structural-metal preparers and erectors	Medium
4227	Survey and market research interviewers	Medium
9613	Sweepers and related labourers	Low
2522	Systems administrators	High
7531	Tailors, dressmakers, furriers, and hatters	Medium
5312	Teachers' aides	Low
2359	Teaching professionals not elsewhere classified	High
2433	Technical and medical sales professionals (excluding ICT)	High
2153	Telecommunications engineers	High
4223	Telephone switchboard operators	Medium
8159	Textile, fur, and leather products machine operators not elsewhere classified	Medium
7516	Tobacco preparers and tobacco products makers	Medium
3324	Trade brokers	High
2643	Translators, interpreters, and other linguists	High
5111	Travel attendants and travel stewards	Low
5113	Travel guides	Low
6112	Tree and shrub crop growers	Low
4131	Typists and word processing operators	Medium
5163	Undertakers and embalmers	Low
7541	Underwater divers	Medium
7534	Upholsterers and related workers	Medium
9122	Vehicle cleaners	Low
3240	Veterinary technicians and assistants	High
5131	Waiters	Low
2513	Web and multimedia developers	High
7212	Welders and flame cutters	Medium
8172	Wood processing plant operators	Medium
7521	Wood treaters	Medium

Source: Author's calculations based on GEIH information, 2016-2018.

Despite information failures in the labour market and their consequences on unemployment and informality rates, countries like Colombia lack a proper labour market information system to identify skill mismatches and employer skill requirements.

The use of online job portals as a potential source of labour market information has gained the attention of researchers and policymakers since these portals can provide quick and relatively low-cost data collection. However, debates continue about the efficacy and robustness of job portals for labour market analysis. Accordingly, this book implements a novel mixed methods approach (e.g. web scraping, text mining, machine learning, etc.) in order to investigate to what extent a web-based model of skill mismatches can be developed for Colombia. This book contributes to our current understanding of the topic by developing a conceptual and methodological approach to identify skills, occupations, and skill mismatches using online job advertisements, which would otherwise be too complex to analyse via other means.







Este libro fue compuesto en caracteres
Amasis 10 puntos e impreso en el año 2020
por Xpress. Estudio Gráfico y Digital SAS,
en Bogotá, D. C., Colombia